Structure Inference for Bayesian Multisensor Scene Understanding

Timothy M. Hospedales and Sethu Vijayakumar

Abstract—We investigate a solution to the problem of multisensor scene understanding by formulating it in the framework of Bayesian model selection and structure inference. Humans robustly associate multimodal data as appropriate, but previous modeling work has focused largely on optimal fusion, leaving segregation unaccounted for and unexploited by machine perception systems. We illustrate a unifying Bayesian solution to multisensory perception and tracking, which accounts for both integration and segregation by explicit probabilistic reasoning about data association in a temporal context. Such an explicit inference of multimodal data association is also of intrinsic interest for higher level understanding of multisensory data. We illustrate this by using a probabilistic implementation of data association in a multiparty audiovisual scenario, where unsupervised learning and structure inference is used to automatically segment, associate, and track individual subjects in audiovisual sequences. Indeed, the structure-inference-based framework introduced in this work provides the theoretical foundation needed to satisfactorily explain many confounding results in human psychophysics experiments involving multimodal cue integration and association.

Index Terms—Sensor fusion, audiovisual, multimodal, detection, tracking, graphical models, model selection, Bayesian inference, speaker association.

1 Introduction

Optimal fusion of redundant multisensor observations has been of much recent interest both for understanding human multimodal perception [10], [2] and for building better machine perception applications [6], [26]. In principle, multisensor fusion is useful to an agent, because more precise inferences about the world can be drawn, given multiple observations with independent noise. The benefit is potentially greater when noise processes in each modality are disparate: Each sensor's strength can potentially compensate for the other's weakness. For example, in humans equipped with auditory and visual senses, vision with high spatial precision can dominate in spatial localization tasks [2], [5], while audition with high temporal precision can dominate in frequency judgment tasks [30], [28].

Research into human behavior has identified various combinations of senses and tasks for which human perceptual sensor fusion (or multimodal integration) is close to the Bayesian optimal. Examples include vision and haptics [10], vision and audition [2], [5], and texture and motion within vision [19]. In the domain of applying statistical techniques to machine perception problems, the fusion of multiple modalities or features is a common technique to improve performance. In speech recognition, for example, visual lip features have been fused with audio data to improve recognition performance [25]. In tracking, performance has been enhanced by the fusion of color, texture, and edge visual features [29], as well as fusing entirely separate audio and video modalities [6], [26], [18], [9].

Notably, all these studies have generally considered 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 will 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 in the same latent source. A more general problem in multisensory 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 long standing interest in the radar community: The association decisions might, for example, be made between a pool of candidate detections and existing aircraft tracks before tracks can be updated on the basis of observations. However, popular methods in this domain [4], [3] have tended to be heuristic heavy due to the strict real-time requirements coupled with typically high-dimensional large data sets, with some notable exceptions [36], [37].

In a probabilistic modeling context, data association is an example of a structure inference or model selection problem. Here, the potential existence of a causal connection between a given pair of latent and observed variables is itself an unknown. Early studies of this type of uncertain structure problem by the probabilistic modeling community described efficient inference for some classes of network using Bayesian Multinets [15]. If potential conditional indepedencies are not known a priori, they can themselves be discovered in data using Context-Specific Independence [8]. The Bayesian Multinet approach has also been applied to infer the (time-varying) connectivity structure in Markov chains [7]. All this is in contrast to *learning* a *fixed* Bayesian network structure from large data sets, which is also topical [32], [24].

• *The authors are with the Institute of Perception, Action, and Behavior, University of Edinburgh, James Clerk Maxwell Bulidng, The King's Building, Mayfield Road, Edinburgh EH9 3JZ. E-mail: {t.hospedales, sethu.vijayakumar}@ed.ac.uk.x*

Manuscript received 24 Feb. 2007; revised 31 Oct. 2007; accepted 26 Dec. 2007; published online 17 Jan. 2008.
Recommended for acceptance by T. Darrell.
For information on obtaining reprints of this article, please send e-mail to: tpami@computer.org, and reference IEEECS Log Number TPAMI-0132-0207.
Digital Object Identifier no. 10.1109/TPAMI.2008.25

Inspired by radar/sonar data association algorithms [12], some machine learning for computer vision applications have begun to consider this issue [27]. Nevertheless, computer vision studies have tended to see data association as a nuisance variable—a prerequisite for correct fusion in a multisensor and/or multitarget context—but otherwise uninteresting and to be integrated out as quickly and efficiently as possible. In contrast, we will argue that for many applications, the association is itself a useful output worthy of careful explicit modeling and consideration. Data association can be of intrinsic interest to understand complex semantic structure in the data. This is clearly the case in problems of audiovisual (AV) perception, where the association represents *who said what*. For example, typical meeting room goals for a human or automatic transcription machine [17] might include understanding speech and identifying the participants. However, without explicitly computing the association between audio and visual observations and latent sources, such an agent might have a notion of "who was there" and "what was said" but not "*who said what*," which is a *relational* concept specifying the existence of causal connections between different variables that is critical to the meeting understanding paradigm. Some recent studies have included the computation of speaker association in AV tracking for meeting analysis using particle filters [14]. To compute data association, an alternative approach to structure inference in explicit parametric models is based on computing and thresholding the mutual information between modalities [34], [11]. However, this has the drawback of being purely a method for estimating association without a principled framework for simultaneous inference of other quantities of interest such as tracking [14] or detection [25] that parametric models can provide.

In this paper, we illustrate a common framework for multisensory perception problems of potentially unknown data association and provide a unifying principled Bayesian account of their solution, reasoning explicitly about the association between observations and latent source states. In Section 2, we introduce the probabilistic foundations of multisensory perception problems where the data association is unknown. To quantifiably highlight the benefits of this method, we describe the task and generic modeling framework by way of a series of toy models. In Section 3, we introduce a probabilistic model capable of representing real AV data of unknown association. While being conceptually the same, the AV model is necessarily significantly more detailed than the generic form. In Section 4, we illustrate results for learning, tracking, and computing association for human targets. We summarize our contributions and their relation to other research in Section 5. Derived learning rules are summarized in Appendix A.

2 Probabilistic Data Association Theory

In order to formalize the perceptual problem of combining information from multiple sensory modalities to obtain an accurate unified percept of the world, we use a probabilistic generative modeling framework. The task of perception can be abstracted to one of performing *inference* in the generative model, where "latent" quantities of interest (for example, the location of a person) and data association (for example, who said what) are both inferred. We can frame such an inference as a model selection (or structure inference) problem, as schematically represented by the graphical models in Fig. 1. Here, observations in two *different modalities* $D = \{x_1, x_2\}$ are potentially generated from a *single* source with latent state l (Fig. 1a). Under this generative model, the source state is assumed to be drawn independently along with binary visibility/occlusion variables $\mathrm{M}_1, \mathrm{M}_2$ in each modality. Subsequently, the observations are 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 (Fig. 1b).

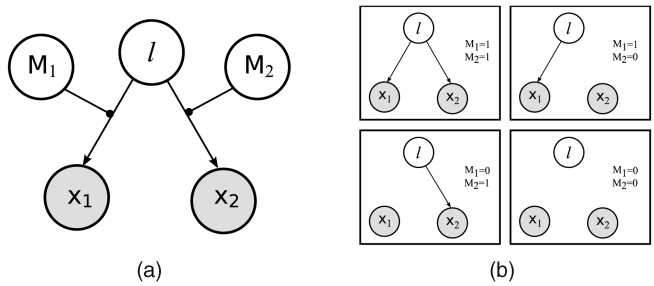


Fig. 1. Graphical models to describe the "unreliable generation" of multimodal observations from a single source. (a) Variable structure interpretation. (b) Variable model interpretation.

Perceptual inference then consists of computing the posterior over the latent state and the generating model (as specified by 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 in) the latent source of interest with probability $p(\mathrm{M}_i = 1|D)$. This will be large if the observation is likely under the foreground distribution (that is, correlated with the prior and other observations) and will be small if it is better explained by the background distribution.

2.1 An Illustrative Example

To illustrate with a toy but concrete example, consider the problem of inferring a one-dimensional latent state l representing a location on the basis of two point observations in separate modalities. For the purpose of this illustration, let the latent location be governed by an informative Gaussian[1] prior $l \sim \mathcal{N}(l|0, p_l)$, with the binomial visibility variables having prior probability $p(\mathrm{M}_i) = \pi_i$ (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 uninformativeness with precision $p_b \to 0$. The joint probability can then be written as

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

If we are purely interested in computing the posterior over the latent state, we integrate over models or structure

1. 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 Section 3.

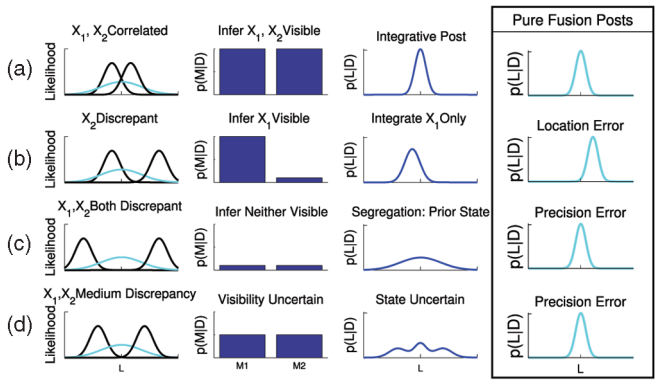


Fig. 2. Schematic of data association inference, given multimodal observations x_i. Likelihoods of the observations in each of the two modalities are in black, and the prior is in gray/cyan. Observations: (a) x_1 and x_2 are strongly correlated. (b) x_2 is strongly discrepant. (c) x_1 and x_2 are both strongly discrepant. (d) x_1 and x_2 are both moderately discrepant.

variables: $\sum_{\mathrm{M}_1,\mathrm{M}_2} p(D, l, \mathrm{M})$. For a 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 the Bayesian Occam's razor [23]. Defining for brevity $m_i \equiv (\mathrm{M}_i = 1)$ and $\overline{m}_i \equiv (\mathrm{M}_i = 0)$, based on (1), we can write down the posteriors as follows:

$$p(\overline{m}_1, \overline{m}_2|D) \propto \mathcal{N}(x_1|0, p_b)\mathcal{N}(x_2|0, p_b), \tag{2}$$

$$p(m_1, \overline{m}_2|D) \propto exp\left(-\frac{1}{2}x_1^2 p_1 p_l/(p_1 + p_l)\right)\mathcal{N}(x_2|0, p_b), \tag{3}$$

$$p(m_1, m_2|D) \propto exp\left[-\frac{1}{2}\frac{x_1^2 p_1(p_2 + p_l) - 2x_1 x_2 p_1 p_2 + x_2^2 p_2(p_1 + p_l)}{p_1 + p_2 + p_l}\right]. \tag{4}$$

The ***structure*** *posterior* $p(\mathrm{M}|D)$ is dependent on the relative data likelihood under the background and the marginal foreground distribution. For example, the posterior of the completely disassociated model (2) 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 associated model (4) depends on the three-way agreement between the observations and the prior. The model structure inference computed using (2)-(4) is plotted as $p(\mathrm{M}|D)$ in Fig. 2 for various illustrative cases. The figure also plots the inferred ***latent state*** *"location" posterior* $p(l|D)$, which can be contrasted with the pure-fusion models (refer to Fig. 2(box)), estimates as follows:

1. *The observations and the prior are all strongly correlated* (Fig. 2a). Both observations are inferred to be associated with the latent source. The location posterior is approximately Gaussian, with $p(l|x_1, x_2) \approx \mathcal{N}(l|\hat{l}, p_{l|x})$, where $p_{l|x} = p_1 + p_2 + p_l$, and $\hat{l} = \frac{p_1 x_1 + p_2 x_2}{p_{l|x}}$. This matches the pure-fusion estimates.
2. *Observation x_2 is strongly discrepant with x_1 and the prior* (Fig. 2b). Sensor 2 is inferred to be occluded. The resulting approximately Gaussian location posterior fuses only x_1 and the prior. $p_{l|x} = p_1 + p_l$, and $\hat{l} = \frac{p_1 x_1}{p_{l|x}}$. Pure-fusion posterior modes can be displaced arbitrarily far from the actual source as a consequence of fusing the unrelated sensor (Fig. 2b (box)).
3. *Observations x_1 and x_2 are strongly discrepant with each other and the prior* (Fig. 2c). Both observations are inferred to be unrelated to the actual source (both sensors occluded), in which case the posterior over the latent state reverts to the prior $p_{l|x} = p_l$, $\hat{l} = 0$. In the pure-fusion models, posterior distributions could indicate dramatically inappropriate overconfidence (Fig. 2c (box)).
4. *The correlation between the observations and the prior is only moderate* (Fig. 2d). The posteriors over structural visibility variables are highly uncertain. The location posterior is a (potentially quad-modal) mixture of Gaussians corresponding to the four possible models. Again, the pure-fusion model displays inappropriate overconfidence over the location (Fig. 2d (box)).

In real-world scenarios, occlusion, sensor failure, or other cause of meaningless observation is almost always possible. In these cases, assuming a typical pure-fusion model (equivalent to constraining $\mathrm{M}_1 = \mathrm{M}_2 = 1$) can result in a dramatically inappropriate inference (as illustrated in Fig. 2 (box) and the explanation above). Examples of these types of effect in real data will be illustrated in Section 4.1. The biggest benefit of this approach, however, will be evident in real-world applications where meaningful sources and observations result in data association (inferred through the structure posterior) having important relational consequence rather than merely ensuring robust tracking.

2.2 Incorporating Temporal Dependencies

To make good use of the techniques described in the Section 2.1, we need appropriate ***prior distributions*** to compute association with and rely upon in the event of complete sensor failure or occlusion. Therefore, for the tracking tasks, we take into account temporal context. In addition to object location, the observation association itself may be correlated in time. For example, if the target passes behind an occluder, it may take some time before it becomes visible again on the other side. To model data with these correlations, we introduce the graphical model in Fig. 3a, in which the state l and model variables M_i are each now connected through time. 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 independent and identically distributed (IID) case. An inference may then consist of computing the posterior over the latent variables at each time t, given all T available observations, $p(l^t, \mathrm{M}^t|x_1^{1:T}, x_2^{1:T})$ (that is, smoothing) if the processing is offline. If the processing must be online, the posterior over the latent variables, given all the data, up to the current time $p(l^t, \mathrm{M}^t|x_1^{1:t}, x_2^{1:t})$ (that is, filtering) may be employed. Multimodal source tracking is performed by computing the posterior of l, marginalizing over possible associations. We have seen previously that the posterior distribution over the location at a given time is potentially non-Gaussian (Fig. 2d).

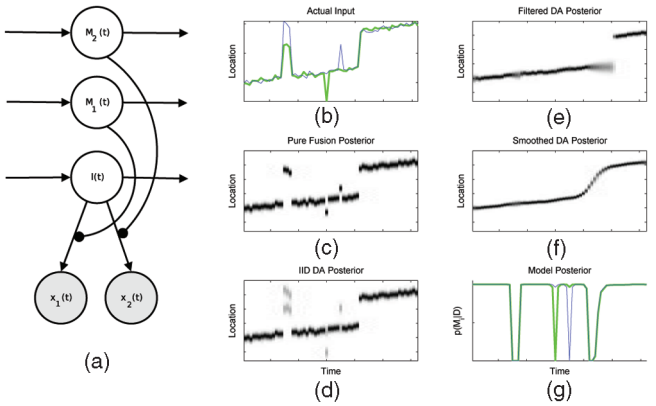


Fig. 3. (a) Graphical model to describe the generation of observations x_i with temporal dependency. (b) Synthetic input data set in modalities x_1 and x_2. Posterior probability of l in (c) pure-fusion model, (d) IID data association model, (e) filtered data association model, and (f) smoothed data association model. (g) Posterior probability of model structure for the temporal data association model.

To represent such general distributions, we can discretize the state space of l, producing a factorial hidden Markov model [16] (FHMM). In this example, an 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 FHMM in terms of the posteriors $\alpha^t \triangleq p(l^t, \mathrm{M}_{1,2}^t|D^{1:t})$ and $\gamma^t \triangleq p(l^t, \mathrm{M}_{1,2}^t|D^{1:T})$:

$$\alpha^t \propto \sum_{l^{t-1},\mathrm{M}_{1,2}^{t-1}} p(D^t|l^t, \mathrm{M}_{1,2}^t) p(l^t|l^{t-1}) \prod_{i=1}^{2} p(\mathrm{M}_i^t|\mathrm{M}_i^{t-1}) \alpha^{t-1}, \quad (5)$$

$$\gamma^t \propto \sum_{l^{t+1},\mathrm{M}_{1,2}^{t+1}} \frac{p(l^{t+1}|l^t) \prod_{i=1}^{2} p(\mathrm{M}_i^t|\mathrm{M}_i^{t-1}) \alpha^t}{\sum_{l^t,\mathrm{M}_{1,2}^t} p(l^{t+1}|l^t) \prod_{i=1}^{2} p(\mathrm{M}_i^t|\mathrm{M}_i^{t-1}) \alpha^t} \gamma^{t+1}. \quad (6)$$

Filtering makes use of the forward α recursion in (5) and smoothing the backward γ recursion in (6), which are analogs of the α and γ recursions in standard HMM inferences. The benefits of temporal context for the inference of the source state and data association are illustrated in Figs. 3b, 3c, 3d, 3e, 3f, and 3g. Fig. 3b illustrates data from a series of T observations, $x_i^t \sim \mathcal{N}(l^t, p)$, $D = \{x_1^t, x_2^t\}_{t=1}^T$, in two independent modalities, of a continuously varying latent source l. These data include some occlusions/sensor failures (where the observations 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. A *pure-fusion model* without temporal context (Fig. 3c) has very limited robustness, as an inference in this model always consists of a simple precision-weighted average over observations. This procedure is not useful, since the disassociated observations can come from an entirely different distribution and hence throw off the average. A *data association model* (Fig. 3d) is slightly more robust, correctly inferring that the pure-fusion generative structure is unlikely when the observations are discrepant. However, without temporal context, it cannot identify which observation was discrepant. Marginalizing over the models, it produces a non-Gaussian multimodal posterior for l. Including some temporal history, an online "*filtered*" *data association model* can infer which observations are discrepant and discount them, producing a much smoother inference (Fig. 3e). In this case, after the discontinuity in state, the fully disassociated observation structure is inferred. Also, based on the temporal diffusion model, an approximately constant location is inferred until enough evidence is accumulated to support the new location. Finally, an offline "*smoothing*" *data association model* (Fig. 3f) infers a robust accurate trajectory. For this case, the marginal posterior of the association variables is shown in Fig. 3g.

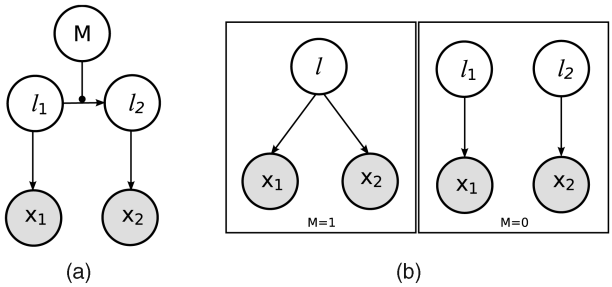


Fig. 4. Graphical models to describe the generation of multimodal observations x_1 and x_2, which may be due to separate sources or one single source. (a) Variable structure representation. (b) Variable model representation.

The illustrative scenarios discussed here generalize in the obvious way to more observations. With many sensors, the disassociation of a small number of discrepant sensors can be inferred, even without prior information. However, in a pure-fusion scheme, even with many sensors, a single highly discrepant sensor can throw off all the others during averaging.

2.3 An Illustrative Example with Multiple Objects

There is another simple way in which two multimodal observations can be generated; that is, each could be generated by a *separate* source instead of a single source. The choice of the multisource versus single-source generating model (Fig. 4b) can also be expressed compactly as a structure inference (Fig. 4a) as before but by using two latent state variables and requiring equality between them if $\mathrm{M} = 1$ and independence if $\mathrm{M} = 0$. 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 ("multiobject semantic"). The occlusion semantic arises, for example, in AV processing, where a source may independently be visible or audible. The multiobject semantic arises, for example, in some psychophysics experiments [30], which will be discussed later.

We will now illustrate the latter case with a toy but concrete example of generating observations in two different modalities x_1 and x_2, which may both be due to a single latent source $(\mathrm{M} = 1)$ or two separate sources $(\mathrm{M} = 0)$. 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 that the prior distributions over the latent locations are Gaussian

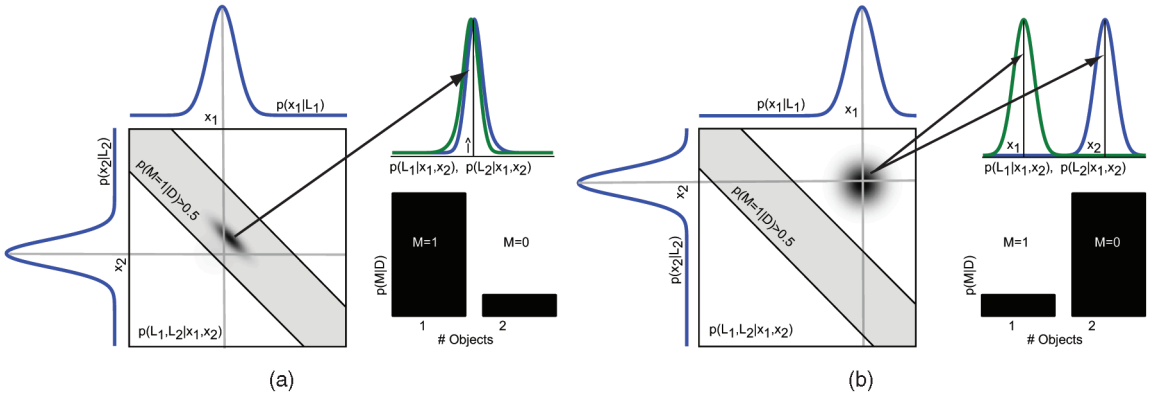


Fig. 5. Inference in multiobject semantic toy model. (a) For correlated inputs $x_1 \approx x_2$, the presence of one objects is inferred, and its location posterior is the probabilistic fusion of the observations. (b) For very discrepant inputs $x_1 \neq x_2$, the presence of two objects is inferred, and the location posterior for each is at the associated observation. (a) Inferring integrative structure from correlated inputs. (b) Inferring segragative from decorrelated inputs.

but tend to uninformativeness. In the multiobject model, the prior over l_is: $p(\mathbf{l}|\mathrm{M}=0) = \mathcal{N}(\mathbf{l}|\mathbf{0}, \mathbf{P}_0)$ 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: $p(\mathbf{l}|\mathrm{M}=1) = \mathcal{N}(\mathbf{l}|\mathbf{0}, \mathbf{P}_1)$ 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 as follows:

$$
\begin{aligned}
p(\mathbf{x}, \mathbf{l}, \mathrm{M}) &= \mathcal{N}(\mathbf{x}|\mathbf{l}, \mathbf{P}_x)\mathcal{N}(\mathbf{l}|\mathbf{0}, \mathbf{P}_0)^{(1-M)}\mathcal{N}(\mathbf{l}|\mathbf{0}, \mathbf{P}_1)^{M}p(\mathrm{M}),\\
p(\mathrm{M}|\mathbf{x}) &\propto \int_{\mathbf{l}} \mathcal{N}(\mathbf{x}|\mathbf{l}, \mathbf{P}_x)\mathcal{N}(\mathbf{l}|\mathbf{0}, \mathbf{P}_0)^{(1-M)}\mathcal{N}(\mathbf{l}|\mathbf{0}, \mathbf{P}_1)^{M}p(\mathrm{M}),\\
p(\mathrm{M}=0|\mathbf{x}) &\propto \mathcal{N}(\mathbf{x}|\mathbf{0}, (\mathbf{P}_x^{-1} + \mathbf{P}_0^{-1})^{-1})p(\mathrm{M}=0),\\
p(\mathrm{M}=1|\mathbf{x}) &\propto \mathcal{N}(\mathbf{x}|\mathbf{0}, (\mathbf{P}_x^{-1} + \mathbf{P}_1^{-1})^{-1})p(\mathrm{M}=1).
\end{aligned}
\tag{7}
$$

A compact representation of the interesting behaviors exhibited is illustrated in Fig. 5. If observations x_1 and x_2 are only slightly discrepant (as depicted by the gray crosshairs in Fig. 5a), 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, that is, $p(\mathbf{l}|\mathbf{x}) \approx \mathcal{N}(\mathbf{l}|\hat{\mathbf{l}}, \mathbf{P}_{l|x})$, where $\hat{\mathbf{l}} = \mathbf{P}_{l|x}^{-1}\mathbf{P}_x\mathbf{x}$, and $\mathbf{P}_{l|x} = \mathbf{P}_x + \mathbf{P}_1$. The location marginals for each l_i are therefore the same and aligned at $\hat{\mathbf{l}}$. However, if x_1 and x_2 are highly discrepant (Fig. 5b), 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, that is, $\hat{\mathbf{l}} = \mathbf{P}_{l|x}^{-1}\mathbf{P}_x\mathbf{x} \approx \mathbf{x}$, and $\mathbf{P}_{l|x} = \mathbf{P}_x + \mathbf{P}_0$.

A real albeit discrete domain in which these multiobject association ideas are relevant are the psychophysics experiments reported in [30] and [31]. In these experiments, a variable number 1-4 of approximately coincident beeps and flashes are played to the subject, who must report how many were actually played on the basis of their noisy sensory input. Since in the real world, events frequently produce correlated multimodal observations, the hypothesis that these observations correspond to the same event(s) is a plausible one for the perceptual system to consider against the hypothesis that they are unrelated. An apparent small discrepancy in the likelihood peaks for beep number and flash number might therefore be more likely to be due to sensory noise in the observation of a single correlated source, whereas an apparent large discrepancy might be more likely to be because the observations are actually unrelated. This integration of similar observations (leading to the flash-beep illusion [30]) and segregation of highly discrepant signals is indeed the observed outcome of these experiments. Using our variable structure interpretation of the problem, we can fit the data reported in [31] exceptionally well within a unified model framework.

2.4 Summary

For the sake of clarity and to highlight key conceptual advantages that our generic modeling framework provides, the observation likelihood models described so far were linear with simple point observations. As we will see in Section 3, observations from real-world data may be high-dimensional and nonlinear and involve extra latents with specific generative model idiosyncrasies. The consequences and benefits for data association, however, are conceptually the same as for this generic framework.

Also, the inferences discussed in this section have been exact. There are various potential approximations such as computing the *location posterior* by using the maximum a posteriori (MAP) model, which may be acceptable but crucially misrepresents the state posterior for regions of the input space with intermediate discrepancy (see Fig. 2d). 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, in turn, depends on both the agreement between the 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 is always selected.

We have illustrated a principled probabilistic framework to address the motivating question of how humans and machines should combine multiple sensing modalities during perception. Many previous probabilistic accounts of human multisensory combination (for example, [10] and [2]) and machine perception systems [6], [26] are special cases of our theory, having explicitly or implicitly assumed a pure-fusion structure. Hence, these do not, for example, exhibit the robust discounting (sensory segregation) of strongly discrepant cues observed in humans [31], [10], but as we have seen, such a segregation is necessary for

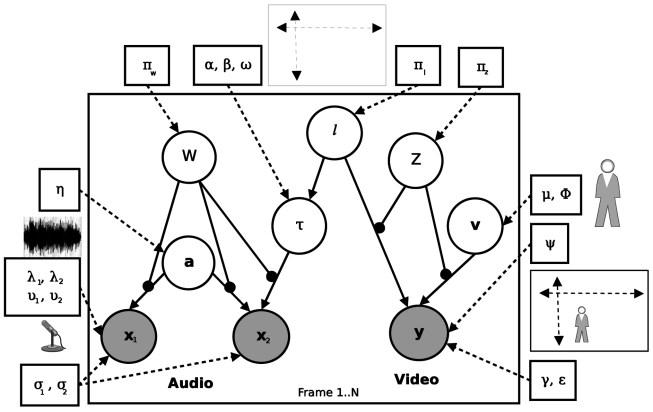


Fig. 6. Graphical model for AV data generation. Refer to Table 1 for the summary of notation.

perception in the real world, since outliers can "break" pure-fusion schemes. The solution to the more general unknown data association problem (in illustrated simple models) has been derived without recourse to heuristics by performing an inference on structure and state variables. This inference depends on the specific parameters of the models, which have so far been assumed to be known. In Section 3, we investigate scaling the model up to include high-dimensional observations, which depend, in complex ways, on the latent state, as well as learning of the models' parameters directly from real sensor data by using the expectation-maximization (EM) algorithm.

3 Bayesian Multisensory Perception for AV Scene Understanding

To illustrate the application of these ideas to a real large-scale machine perception problem, we consider a model and task inspired by [6]: that of unsupervised learning and inference with AV input. Beal et al. [6] demonstrated the inference of an AV source location and unsupervised 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 underlying machinery in [6] is itself largely based on the *transformed mixture of Gaussians* (TMG) framework [13], which, as we shall see, allows efficient inference.

The AV localization part of the task is similar to the task in psychophysics experiments such as [2], where humans are 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 and temporal independence were previously assumed), that is, whether the source should be associated with both modalities or only one or if there is no source present at all.

3.1 Generative Model

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. 6, along with Table 1, which summarizes the notation. 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 and due to the nature of our data, we only consider source translation along the azimuth, so l effectively ranges over all the x-axis pixels of the image.[2] This could easily be expanded to include y-axis translation, as in [6], without a significant computational cost. First, consider the all-visible (pure-fusion) case $(\mathrm{W}, \mathrm{Z} = 1)$. The video appearance $\mathbf{v}$ is sampled from a diagonal Gaussian distribution $\mathcal{N}(\mathbf{v}|\mu\mu, \phi)$, with parameters defining its soft template. The observed video pixels are generated by sampling from the uniform diagonal 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 precision Gaussian, that is, $\mathcal{N}(\mathbf{a}|\mathbf{0}, \eta\mathbf{I})$. (This model can potentially use a Toeplitz matrix η to represent a spectral structure in the signal [6], but for simplicity, we consider the uniform diagonal case here.) The time delay τ between the signals at each microphone is drawn as a linear function of the translation of the source $\mathcal{N}(\tau|\alpha l + \beta, \omega)$. 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 τ samples relative to $\mathbf{x}_1$, that is, $\mathbf{x_1} \sim \mathcal{N}(\mathbf{x}_1|\mathbf{a}, \upsilon_1\mathbf{I})$, and $\mathbf{x_2} \sim \mathcal{N}(\mathbf{x}_2|\mathbf{T}_\tau\mathbf{a}, \upsilon_2\mathbf{I})$, respectively. If the video modality is occluded $(\mathrm{Z} = 0)$, the observed video pixels are drawn from a very generic Gaussian background distribution $\mathcal{N}(\mathbf{y}|\gamma\mathbf{1}, \epsilon\mathbf{I})$, independent of l and the audio data. If the audio modality is silent $(\mathrm{W} = 0)$, the samples at each speaker are drawn from very generic background distributions $\mathcal{N}(\mathbf{x}_i|\mathbf{0}, \sigma_i\mathbf{I})$, independent of each other, l, and the video.

To describe the generation of a series of correlated frames, the IID observation model in Fig. 6 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 to Fig. 3a). The state evolution distribution over the location shift is defined in the standard way $p(l^{t+1}|l^t) = \Gamma_{[l^t, l^{t+1}]}$, where the subscripts pick out the appropriate element of the matrix Γ. The observability transitions are defined similarly as $p(\mathrm{W}^{t+1}|\mathrm{W}^t) = \Theta_{[w^t, w^{t+1}]}$ and $p(\mathrm{Z}^{t+1}|\mathrm{Z}^t) = \Omega_{[z^t, z^{t+1}]}$. Suppressing unambiguous indexing by t for clarity, the joint probability of the model, including all visible $D = \{\mathbf{x}_1, \mathbf{x}_2, \mathbf{y}\}_{t=1}^T$ and hidden variables $H = \{\mathbf{a}, \mathbf{v}, \tau, l, \mathrm{W}, \mathrm{Z}\}_{t=1}^T$, given all the parameters

$$\theta = \{\lambda_{1,2}, \upsilon_{1,2}, \eta, \alpha, \beta, \omega, \pi_l, \mu, \phi, \Psi, \Gamma, \Theta, \Omega, \pi, \pi_z, \gamma, \epsilon, \sigma_{1,2}\},$$

factorizes as

$$\begin{aligned} p(D, H|\theta) = & \prod_{t=0}^{T-1} p(l^{t+1}|l^t) p(\mathrm{W}^{t+1}|\mathrm{W}^t) p(\mathrm{Z}^{t+1}|\mathrm{Z}^t) \\ & \cdot \prod_{t=1}^{T} p(\mathbf{x}_1|\mathrm{W}, \mathbf{a}) p(\mathbf{x}_2|\mathrm{W}, \mathbf{a}, \tau) p(\mathbf{a}) p(\tau|l) p(\mathbf{v}) p(\mathbf{y}|\mathrm{Z}, \mathbf{v}, l), \\ & \Bigg(\prod_{t=1}^{T} \mathcal{N}(\mathbf{x}_1|\mathbf{a}, \upsilon_1)^w \mathcal{N}(\mathbf{x}_1|\mathbf{0}, \sigma_1)^{\overline{w}} \mathcal{N}(\mathbf{a}|0, \eta) \\ & \cdot \mathcal{N}(\mathbf{x}_2|\mathbf{T}_\tau\mathbf{a}, \upsilon_2)^w \mathcal{N}(\mathbf{x}_2|\mathbf{0}, \sigma_2)^{\overline{w}} \mathcal{N}(\tau|\alpha l + \beta, \omega) \\ & \cdot \mathcal{N}(\mathbf{y}|\mathbf{T}_l\mathbf{v}, \Psi\mathbf{I})^z \mathcal{N}(\mathbf{y}|\gamma\mathbf{1}, \epsilon\mathbf{I})^{\overline{z}} \mathcal{N}(\mathbf{v}|\mu, \phi) \Bigg) \\ & \cdot \prod_{t=0}^{T-1} \Gamma_{l^t, l^{t+1}} \Theta_{w^t, w^{t+1}} \Omega_{z^t, z^{t+1}} \cdot \pi_l \pi_z \pi_w. \end{aligned} \tag{8}$$

2. As l is actually a translation, its range during tracking can be constrained to the region around the current location for computational efficiency [21]; however, we need not do this.

TABLE 1
Summary of AV Model Variables and Parameters

Variables		Parameters	
$\mathbf{x}_1, \mathbf{x}_2$	Observed signal at microphones	υ_1, υ_2	Precision of microphone speech reception
		λ_1, λ_2	Microphone speech attenuation
		$\sigma_{1,2}$	Precision of background noise
$\mathbf{a}$	Speech signal	η	Precision of speech signal
τ	Inter microphone time delay	α, β, ω	Determine time delay as linear function of source location
W	Audio association	π_w, Θ	Prior and dynamic probability of audio association
$\mathbf{y}$	Observed video frame	Ψ	Foreground precision of video camera
		γ, ϵ	Video background mean and precision
$\mathbf{v}$	Video appearance	μ, ϕ	Mean and precision of foreground video appearance
Z	Video association	π_z, Ω	Prior and dynamic probability of video association
l	Discrete source location (pixels)	π_l, Γ	Prior and dynamic probability of source location

For convenient reference, all of the variables and parameters used in the model are summarized in Table 1.

3.2 Inference

Let us first consider an inference, given a single frame of data. The Bayesian network described so far gives us the joint probability in (8). Due to the structure of the model, the full posterior for over all the latent variables for a single frame of data factors into independently computable terms:

$$p(\mathbf{a}, \mathbf{v}, \tau, l, \mathrm{W}, \mathrm{Z}|\mathbf{x}_1, \mathbf{x}_2, \mathbf{y}) = p(\mathbf{a}|\tau, \mathrm{W}, D)p(\mathbf{v}|l, \mathrm{Z}, D) \cdot p(\tau|l, \mathrm{W}, D)p(l, \mathrm{W}, \mathrm{Z}|D). \quad (9)$$

The quantities of ultimate interest for this AV scene understanding task are the ***location of the source*** and its ***visibility and audibility***. The posterior over these quantities is contained in the factor $p(l, \mathrm{W}, \mathrm{Z}|D)$, which is all that need to be computed for an efficient performance once the model is trained. However, during the training phase, it will be necessary to compute each component of the full posterior for learning with the EM algorithm. The factors $p(\mathbf{a}|\tau, \mathrm{W}, D)$ and $p(\mathbf{v}|l, \mathrm{Z}, D)$ are the distributions over the latent signals before noise and will be used to train the audio and video recognition models, respectively, during the M step. The audio signal posterior could also serve as the input to any other downstream audio processing, for example, speech recognition. Finally, the joint posterior over the time delay and location is contained in the product $p(\tau, l, \mathrm{W}, \mathrm{Z}|D) = p(\tau|l, \mathrm{W}, \mathrm{Z}, D)p(l, \mathrm{W}, \mathrm{Z}|D)$, which will be used to train the AV link parameters $\{\alpha, \beta, \omega\}$.

3.2.1 Latent Signal Posteriors

In this section, we derive the posteriors over the latent variables, which will be necessary for training the models of the audio and video signals, as well as the AV-link parameters. These are all conditioned on the location l and observability $z \equiv (Z = 1)$ or $w \equiv (W = 1)$ in the relevant modality.

The joint distribution over the current video data $\mathbf{y}$ and appearance $\mathbf{v}$ is the product of Gaussians $p(\mathbf{y}, \mathbf{v}|l, z) = p(\mathbf{y}|\mathbf{v}, l, z)p(\mathbf{v})$ and is hence also Gaussian. Conditioning on the data $\mathbf{y}$, the posterior $p(\mathbf{v}|l, z, \mathbf{y})$ of the current video appearance is Gaussian, with statistics $p(\mathbf{v}|l, z, \mathbf{y}) = \mathcal{N}(\mathbf{v}|\mu_{\mathbf{v}|\mathbf{y},l,z}, \nu_{\mathbf{v}|z})$, where

$$\mu_{\mathbf{v}|\mathbf{y},l,z} = \nu_{\mathbf{v}}^{-1}(\phi\mu + \mathbf{T}_l^T\Psi\mathbf{y}), \quad (10)$$

$$\nu_{\mathbf{v}|z} = \phi + \Psi. \quad (11)$$

$p(\mathbf{v}|l, z, \mathbf{y})$ is the inference about the source's appearance before being corrupted by noise Ψ and translation $\mathbf{T}_l$. For the purpose of video enhancement, the mean $\mu_{\mathbf{v}|\mathbf{y},l,z}$ of this distribution can be interpreted as the denoised estimate of the image, with foreground obstructions removed [13]. It is therefore intuitive that this should be used to train the video appearance parameters (μ, ϕ) during learning.

Similar to the structure for video, the joint distribution over the current audio data $\mathbf{x}_1$, $\mathbf{x}_2$, and latent signal $\mathbf{a}$ is the product of Gaussians

$$p(\mathbf{x}_1, \mathbf{x}_2, \mathbf{a}|\tau, w) = p(\mathbf{x}_1|\mathbf{a}, w)p(\mathbf{x}_2|\mathbf{a}, \tau, w)p(\mathbf{a}).$$

Conditioning on the data $\mathbf{x}_1$, $\mathbf{x}_2$, the posterior $p(\mathbf{a}|\tau, w, \mathbf{x}_1, \mathbf{x}_2)$ is Gaussian, with statistics $p(\mathbf{a}|\mathbf{x}, \tau, w) = \mathcal{N}(\mathbf{a}|\mu_{\mathbf{a}|\mathbf{x},\tau,w}, \nu_{\mathbf{a}|w})$, where

$$\mu_{\mathbf{a}|\mathbf{x},\tau,w} = \nu_{\mathbf{a}}^{-1}(\lambda_1\upsilon_1\mathbf{x}_1 + \lambda_2\upsilon_2\mathbf{T}_\tau^T\mathbf{x}_2), \quad (12)$$

$$\nu_{\mathbf{a}|w} = \eta + \lambda_1^2\upsilon_1 + \lambda_2^2\upsilon_2. \quad (13)$$

The mean $\mu_{\mathbf{a}|\mathbf{x},\tau,w}$ represents the best estimate for the true speech signal.

The posterior $p(\tau|w, l, \mathbf{x}_1, \mathbf{x}_2)$ over the interaural time delay τ is a discrete distribution, which turns out to be closely related to the cross correlation between the signals. It can be derived in terms of the audio parameters λ_1, λ_2, υ_1, and υ_2, the generative model $p(\tau|l) = \mathcal{N}(\tau|\alpha l + \beta, \omega)$, and the sufficient statistic $\nu_{\mathbf{a}|w}$ as follows:

$$p(\tau|w, l, \mathbf{x}_1, \mathbf{x}_2) = \frac{\int_{\mathbf{a}} p(\mathbf{x}_1|\mathbf{a}, w)p(\mathbf{x}_2|\mathbf{a}, \tau, w)p(\mathbf{a})p(\tau|l)\,}{p(\mathbf{x}_1, \mathbf{x}_2|l, w)}, \quad (14)$$

$$\propto p(\tau|l)\exp(\lambda_1, \lambda_2, \upsilon_1, \upsilon_2, c_\tau), \quad (15)$$

$$c_t = \nu_{\mathbf{a}|w}^{-1}\sum_i x_{[1,i-\tau]}x_{[2,i]}. \quad (16)$$

3.2.2 Marginal Observation Likelihoods

The marginal observation likelihoods for each modality, that is, video $p(\mathbf{y}|\mathrm{Z}, l)$ and audio $p(\mathbf{x}_1, \mathbf{x}_2|\mathrm{W}, l)$, will prove convenient to have at hand when computing the final posterior quantity $p(l, \mathrm{W}, \mathrm{Z}|D)$. We therefore derive them in this section. These likelihoods are useful for thinking about data association and are exactly analogous to the observation likelihoods introduced in Section 2.

For a visible target z, the marginal video likelihood $p(\mathbf{y}|z, l)$ is again derived from the jointly Gaussian $p(\mathbf{y}, \mathbf{v}|z, l) = p(\mathbf{y}|\mathbf{v}, l, z)p(\mathbf{v})$. Integrating out the video appearance $\mathbf{v}$, we have $p(\mathbf{y}|z, l) = \int_{\mathbf{v}} p(\mathbf{y}, \mathbf{v}|z, l)$, which is Gaussian, with statistics $p(\mathbf{y}|z, l) = \mathcal{N}(\mathbf{y}|\mu_{\mathbf{y}|l,z}, \nu_{\mathbf{y}|l,z})$, where

$$\mu_{\mathbf{y}|l,z} = \mathbf{T}_l \mu, \tag{17}$$

$$\nu_{\mathbf{y}|l,z} = (\Psi^{-1} + \mathbf{T}_l \phi^{-1} \mathbf{T}_l^T)^{-1}. \tag{18}$$

Video disassociation $\overline{z}$ could be due to various causes, including the absence of the target, occlusion by another object or sensor failure. The likelihood of the data, given $\overline{z}$, is therefore defined by a very general background distribution:

$$p(\mathbf{y}|l, \overline{z}) = \mathcal{N}(\mathbf{y}|\gamma \mathbf{1}, \epsilon \mathbf{I}). \tag{19}$$

Note that this is now independent of the location l. For the background video distribution, a diagonal Gaussian is also possible such as in the TMG formulation [13], but the more generic uniform diagonal Gaussian will turn out to be more useful in Section 4.3.

For an audible $w \equiv (\mathrm{W} = 1)$ target, the marginal likelihood $p(\mathbf{x}_1, \mathbf{x}_2|\tau, w)$ is also derived from the jointly Gaussian $p(\mathbf{x}_1, \mathbf{x}_2, \mathbf{a}|\tau, w) = p(\mathbf{x}_1|\mathbf{a}, w)p(\mathbf{x}_2|\mathbf{a}, \tau, w)p(\mathbf{a})$. Integrating out the audio signal $\mathbf{a}$, we have $p(\mathbf{x}_1, \mathbf{x}_2|\tau, w) = \int_{\mathbf{a}} p(\mathbf{x}_1, \mathbf{x}_2, \mathbf{a}|\tau, w)$, which is given by

$$p(\mathbf{x}_1, \mathbf{x}_2|\tau, w) \propto K \exp(\mu_{\mathbf{a}|t,\mathbf{x},w}^T \nu_{\mathbf{a}|w} \mu_{\mathbf{a}|t,\mathbf{x},w}), \tag{20}$$

where K contains terms not dependent on τ. We are, however, ultimately interested in the marginal likelihood, given the *location*, $p(\mathbf{x}_1, \mathbf{x}_2|l, w)$. To obtain this from (20), we combine it with the posterior over the discrete τ (computed in (15)) and numerically integrate τ out (see (21)). Similar to the video model, the marginal likelihood for background noise, conditioned on audio disassociation $\overline{w}$, is a simple background distribution, independent of l (see (22)):

$$p(\mathbf{x}_1, \mathbf{x}_1|l, w) = \sum_{\tau} p(\mathbf{x}_1, \mathbf{x}_2|\tau, w)p(\tau|w, l, \mathbf{x}_1, \mathbf{x}_2), \tag{21}$$

$$p(\mathbf{x}_1, \mathbf{x}_1|l, \overline{w}) = \mathcal{N}(\mathbf{x}_1|\mathbf{0}, \sigma_1 \mathbf{I})\mathcal{N}(\mathbf{x}_2|\mathbf{0}, \sigma_2 \mathbf{I}). \tag{22}$$

Note that the background audio likelihood has eliminated the *intramodality* correlation between $\mathbf{x}_1$ and $\mathbf{x}_2$. In an alternate formulation of the audio background distribution, conditioning on disassociation $\overline{w}$ could simply eliminate the *intermodality* correlation (that is, by making $p(\tau|l, \overline{w})$ uniform instead of peaked) and index a new precision $\eta_{\overline{w}}$ instead of eliminating the intramicrophone correlation entirely. We will make use of this in Section 4.3.

3.2.3 Location and Association Posterior

We can now relate detection and tracking in the more complex AV probabilistic model to the generic cases discussed in Section 2. For a single frame, the quantity of interest for this task is that of audibility, visibility, and location, given the data $p(\mathrm{W}, \mathrm{Z}, l|D)$. This is analogous to the posterior over the model and location $p(\mathrm{M}, l|D)$ discussed in the generic case (1). With the AV marginal likelihoods (17), (18), (19), (20), (21), and (22), as computed in Section 3.2.2, we can also compute $p(\mathrm{W}, \mathrm{Z}, l|D)$ analogously as

$$p(\mathrm{W}, \mathrm{Z}, l|D) = p(\mathbf{y}|\mathrm{Z}, l)p(\mathbf{x}_1, \mathbf{x}_1|l, \mathrm{W})p(\mathrm{Z})p(\mathrm{W})p(l), \tag{23}$$

$$\begin{aligned} &= \mathcal{N}(\mathbf{y}|\mu_{\mathbf{y}|l,z}, \nu_{\mathbf{y}|l,z})^z \mathcal{N}(\mathbf{y}|\gamma \mathbf{1}, \epsilon \mathbf{I})^{\overline{z}} \\ &\cdot \left(\sum_t p(\mathbf{x}_1, \mathbf{x}_2|\tau, w)p(\tau|w, l) \right)^w \\ &\cdot (\mathcal{N}(\mathbf{x}_1|\mathbf{0}, \sigma_1 \mathbf{I})\mathcal{N}(\mathbf{x}_2|\mathbf{0}, \sigma_2 \mathbf{I}))^{\overline{w}} p(Z)p(W)p(l)/p(D). \end{aligned} \tag{24}$$

When computing the filtered or smoothed posterior from multiple frames in the toy model, we used the individual observations with the FHMM recursions (5) and (6). In the AV case, the filtered or smoothed posterior $p(\mathrm{W}^t, \mathrm{Z}^t, l^t|D^{1:T})$ is computed analogously by using new marginal likelihoods $p(\mathbf{y}|l, \mathrm{Z})$ and $p(\mathbf{x}_1, \mathbf{x}_1|l, \mathrm{W})$ in recursions (5) and (6).

3.3 Learning

All the parameters in this model

$$\theta = \{\lambda_{1,2}, \upsilon_{1,2}, \eta, \alpha, \beta, \omega, \pi_l, \mu, \phi, \Psi, \Gamma, \Theta, \Omega, \pi_w, \pi_z, \gamma, \epsilon, \sigma_{1,2}\}$$

are jointly optimized by a standard EM procedure. The inference of the posterior distribution over hidden variables H, given the observed data D, $p(H|D)$ (as computed in (9)) is alternated with the optimization of the expected complete log likelihood or free energy with respect to the parameters: $\frac{\partial}{\partial \theta} \int_H p(H|D) lg \frac{p(H,D)}{p(H|D)}$.

The update for the mean μ of the source visual appearance distribution is given by

$$\mu \leftarrow \sum_{t,l} p(l^t, z^t|D^{1:T}) \mu_{\mathbf{v}|\mathbf{y},l,z}^t / \sum_t p(z^t|D^{1:T}). \tag{25}$$

This is defined in terms of the posterior mean $\mu_{\mathbf{v}|\mathbf{y},l,z}^t$ of the video appearance, given the data D, for each frame t and translation l, as inferred during the E step in (10). 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 of the background noise is given by

$$\sigma_i^{-1} \leftarrow \sum_t p(\overline{w}^t|D^{1:T})(\mathbf{x}_i^t)^T \mathbf{x}_i^t / N_f \sum_t p(\overline{w}^t|D^{1:T}), \tag{26}$$

where N_f specifies the number of samples per audio frame. Again, it is intuitive that the estimate of the background variance should be a weighted sum of the square of signals at each frame, where the weighting is the posterior probability that the source was silent in that frame. In an IID context, the posterior over the relevant variables l^t and W^t is only dependent on the current observation D^t, so the marginals $p(w^t|D^t)$, etc., would replace those used above for weighting. A full list of updates is given in Appendix A.

3.4 Computational and Implementation Details

In Sections 3.4.1 and 3.4.2, we detail how we can improve the efficiency of the computationally expensive steps in

inference and learning and how we can ensure numerical stability during EM convergence.

3.4.1 Efficiency

The major computationally intensive steps in the inference are the computation of the observation likelihoods for *every single discrete position* l (17) or *time delay* τ (20) and the computation of the posterior over τ (15). Upon closer inspection, these equations can be reexpressed in terms of correlations and convolutions allowing efficient computation by a fast Fourier transform (FFT). For example, considering the posterior over τ, from (15), we have

$$\log p(\tau|w,l,\mathbf{x}_1,\mathbf{x}_2) = \\ = \lambda_1\lambda_2 v_1 v_2 \nu_{\mathbf{a}}^{-1} \sum_i \mathbf{x}_1[i-\tau]\mathbf{x}_2[i] + K_1, \quad (27)$$

$$= K_2 \mathrm{Corr}(\mathbf{x}_1,\mathbf{x}_2) + K_1. \quad (28)$$

Considering the audio likelihood $p(\mathbf{x}_1,\mathbf{x}_2|\tau,w)$, from (20), we have

$$\log p(\mathbf{x}_1,\mathbf{x}_2|\tau,w) = \mu_{\mathbf{a}|t,\mathbf{x},w}^T \nu_{\mathbf{a}|w} \mu_{\mathbf{a}|t,\mathbf{x},w} + K_1, \\ = v_{\mathbf{a}|w}^{-1} \sum_i \left(\lambda_1^2 v_1^2 \mathbf{x}_1[i]^2 + 2\lambda_1\lambda_2 v_1 v_2 \mathbf{x}_1[i]\mathbf{x}_2[i+\tau] \right. \\ \left. + \lambda_2^2 v_2^2 \mathbf{x}_2[i]^2\right) + K_1, \quad (29)$$

$$= v_{\mathbf{a}|w}^{-1} \sum_i \left(\lambda_1^2 v_1^2 \mathbf{x}_1[i]^2 + 2\lambda_1\lambda_2 v_1 v_2 \mathrm{Corr}(\mathbf{x}_1,\mathbf{x}_2) \right. \\ \left. + \lambda_2^2 v_2^2 \mathbf{x}_2[i]^2\right) + K_1. \quad (30)$$

The expensive quadratic terms involving both i and τ in (27) and (29) have been expressed as an efficiently computable correlation in (28) and (30).

The learning procedure also involves many potentially computationally expensive steps. For example, (25) requires saving the inferred appearance vector $\mu_{\mathbf{v}|l,\mathbf{y},z}$ for every possible discrete position l and then computing their weighted sum. Computing, even storing, $\mu_{\mathbf{v}|l,\mathbf{y},z}$ for all locations l is potentially expensive for larger images $\mathbf{y}$. Reexpressing the update directly in terms of the convolutions of data rather than the inference result $\mu_{\mathbf{v}|l,\mathbf{y},z}$ allows for both space-efficient and time-efficient inference (see [13] for details).

3.4.2 Numerical Stability

There is one major numerical issue in the algorithm as described so far. We can compute the log-likelihoods $\log p(\mathbf{y}|l,\mathrm{Z})$ for individual values of Z. However, during early cycles of learning, before the parameters are well refined, the likelihood of one model may be much greater than the other such that the likelihood $p(\mathbf{y}|l,\mathrm{Z})$ and, hence, the posterior $p(\mathrm{Z},l|\mathbf{y}) \propto p(\mathbf{y}|l,\mathrm{Z})p(l)p(\mathrm{Z})$ are in danger of underflow for one or other values of Z. Constraining entries in the table $p(\mathrm{Z},l|\mathbf{y})$ to be above a minimum small value during normalization is insufficient. For example, information about the *shape* of the associated log likelihood $\log p(\mathbf{y}|l,z)$ as a function of l is still potentially lost if $p(\mathbf{y}|l,z) \ll p(\mathbf{y}|\overline{z})$. This shape information is important for updates such as (25), which are necessary to properly refine the templates.

To help EM converge in a numerically stable way, for the first few cycles, we therefore modify the computation of log-likelihoods in the E step to constrain the less likely model to be at most K less likely than the other. That is, if, for example, $\log p(\mathbf{y}|l,\overline{z}) > K + \log p(\mathbf{y}|l,z)$, then $\log p(\mathbf{y}|l,z)$ is replaced with $logp(\mathbf{y}|l,\overline{z}) - \mathrm{argmax}_l\{\log p(\mathbf{y}|l,z)\} - K + \log p(\mathbf{y}|l,z)$. That is, $p(\mathbf{y}|l,z)$ is not allowed to be more than $\exp(K)$ less likely than $p(\mathbf{y}|l,\overline{z})$. Values of about $K = 10$ seem to be suitable. The same procedure is performed for the audio likelihoods $p(\mathbf{x}_1,\mathbf{x}_2|\tau,\mathrm{W})$.

4 Robust AV Scene Understanding

In this section, we will present results for unsupervised learning and inference in the model presented in Section 3 using real-world raw AV data. The inference of the posterior $p(l^t,\mathrm{W}^t,\mathrm{Z}^t|D)$ corresponds to source detection via W and Z, source tracking via l, and AV source verification if $w \wedge z$. The unsupervised learning of the video parameters (μ,ϕ) corresponds to learning a soft visual template for the object to be tracked. This is in contrast to many other tracking techniques, which require operator specification of the object to be tracked. Moreover, many other AV multimodal systems require careful calibration of the microphone and camera parameters. In this model, these parameters are encompassed by the model AV link parameters (α,β,ω), which are also learned, rendering the model self calibrating.

4.1 Inferring the Behavior of an AV Source: Detailed Example

Results for an illustrative AV sequence after 25 cycles of EM are illustrated in Fig. 7. In this sequence, the user is initially walking and talking, is then occluded behind another person while continuing to speak, and then continues walking while remaining silent. Fig. 7a illustrates three representative video frames from each of these segments, with the inferred data association and location superimposed. In Figs. 7c-7f, the performance of different variants of the tracking algorithm on this data set are compared. Likelihood and posterior modes, rather than full location distributions, are shown for clarity. Audio and video likelihood peaks are indicated by circles and triangles, respectively. The intrinsic imprecision in the audio likelihood compared to that of the video is clear in their relative spread. In each case, the mode of the final location posterior is indicated by the continuous line.

4.1.1 Tracking with IID Pure-Fusion Model

In the simplest IID pure-fusion model, we constrain $\mathrm{W},\mathrm{Z} = 1$ and use the prior π_l instead of the transition matrix Γ. Notice that this now corresponds to the original model of Beal et al. [6]. The location inference is correct, where the multimodal observations are indeed associated (Fig. 7c). The video modality dominates the fusion, as it has much higher precision (that is, the likelihood function is much sharper), and the posterior is still therefore correct during the visible but silent period, where the weaker peaks in the audio likelihood are spurious. While the person in the video foreground is occluded but speaking, the audio likelihood peaks are generally appropriately clustered. However, the next best match to the learned dark foreground template usually happens to be the filing cabinet in one corner or monitor in the other. With pure fusion, the incorrect but still

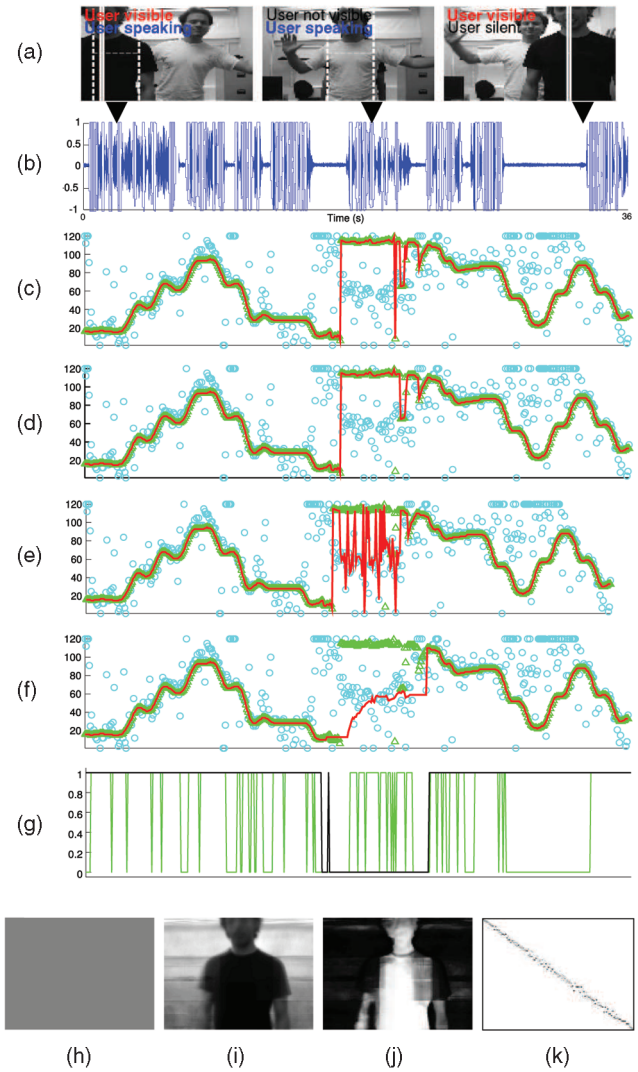


Fig. 7. AV data association and inference results. (a) and (b) Video samples and audio data from a sequence, respectively, where the user is first visibly walking and speaking, is then occluded but still speaking, and is finally visible and walking but silent. (c) Inferred MAP location with IID pure-fusion model. (d) Filtered tracking pure-fusion model. (e) IID data association model. (f) Filtered tracking data association model. Inference based on audio observation alone is shown in circles, video observation alone in triangles, and combined inference by the full (red) line. (g) Posterior probability of visibility (black) and audibility (gray/green) during the sequence. (h) Initial video appearance after learning. (i) and (j) Final video appearance after learning. (k) Final location state-transition matrix after learning.

relatively precise video likelihood dominates the less precise audio likelihood, resulting in a wildly inappropriate posterior.

4.1.2 Tracking with Filtered Pure-Fusion Model

In this case also, we constrain W, $Z = 1$ but now enable temporal tracking with the transition matrix Γ. This is analogous to the multiobservation Kalman filter, which is a standard technique for multimodal tracking. Here, a similar type of error, as described in the pure-fusion case, is made when filtered tracking is used (Fig. 7d). The only difference is that because of the tracking functionality, the jump between the two incorrect locations is eliminated, and the more common of the two previous erroneous locations is focused on.

4.1.3 Tracking with IID Data Association Model

In the IID data association model (Fig. 7e), we do not consider temporal tracking, but we infer W and Z and marginalize over them for localization. The video modality is correctly inferred with high confidence to be disassociated during the occluded period, because the template match is poor. The final posterior during this period 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.

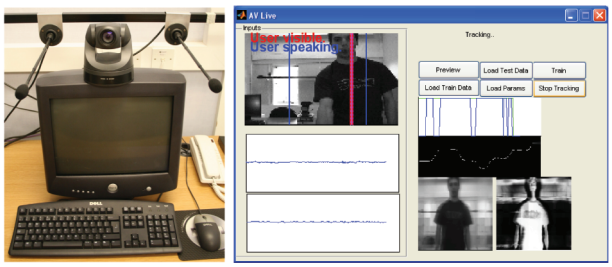
4.1.4 Tracking with Filtered Data Association Model

In the full data association tracking model, we compute the full posterior $p(l^t, \mathrm{W}^t, \mathrm{Z}^t | D^{1:t})$ at every time t. The data association posterior $p(\mathrm{W}^t, \mathrm{Z}^t | D^{1:t})$ (Fig. 7g) correctly represents the visibility and audibility of the target at the appropriate times, and the information from each of the sensors is appropriately weighted for localization. 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. 7f). The inferred data association (Fig. 7g) is used to label the frames in (Fig. 7a) with the user's speaking/visibility status. To cope with intermittent cues, previous multimodal machine perception systems in this context have relied on the observations of discrepant modalities providing uninformative likelihoods [26], [6]. This may not always be the case (see Fig. 7d), as evident from our example video sequence, where only the data association models succeed during the video occlusion.

Using 120×100 pixel video frames and 1,000 sample audio frames, our Matlab implementation can perform online real-time (filtered) tracking at 50 fps after learning, which proceeds at 10 fps. To use our system, the user approaches an AV-equipped PC (Fig. 8a) and presses the train button on our application interface (Fig. 8b), after which he/she is requested to intermittently move around and speak while 20 s of training data is collected. After data collection, the EM algorithm is initiated, and training takes about 5 minutes. Once trained, the user is subsequently AV detected, verified, and tracked in real-time. If the same person will reuse the system, the parameters of the Bayesian network can be saved and reloaded later to avoid retraining.

4.2 Inferring the Behavior of an AV Source: Quantitative Evaluation

In Section 4.1, we described in detail the processing of an example sequence, which illustrated most of the important qualitative differences in the behavior between the model variants. In this section, we describe the results of a more extensive quantitative evaluation of the models against the ground truth for a variety of sequences. Rather than using the typical manual markup of video sequence ground truth, we chose to apply the promising but underexplored approach of mechanical generation of test data.



(a) (b)

Fig. 8. (a) Computer equipped with camera and microphone pair. (b) User interface for unsupervised appearance learning, AV data association, and tracking.

Fig. 9. Sound positioning device. Position is controllable across 2.2 m in the horizontal plane to a 1-mm accuracy. Inset: the speaker generating the audio source.

4.2.1 Evaluation Procedure

We constructed a computer-positionable AV source using an off-the-shelf speaker component driven on a rail by stepper motor (Fig. 9). This allowed us to control precisely and repeatably the source location along 2.2 m of the horizontal plane and to control its audibility and visibility, as required for the evaluation of AV tracking. This automatic generation of training data provided us with a source of ground-truth information, without the need for manual labeling. Different visual appearances could be selected by attaching different objects to the movable source carriage.

To evaluate the models' performance in a variety of different conditions, we controlled four separate variables, for a total of 24 different conditions, as follows: When present, the audio signal was played at either high or low volume and was composed of low-pass filtered noise below 16,384, 2,048, or 256 Hz (making audio localization increasingly imprecise). Although somewhat less realistic, a simple noise signal was used, rather than a recorded speech, so that ground-truth audibility could be clearly controlled without the uncertainty in labeling of interword pauses, etc. [33]. Indeed, this has resulted in the audio-only tracking result to have significantly less variance than, for example, the speech-based example illustrated in Fig. 7. The visual appearance of a person was simulated by attaching two possible different sets of clothing, and the room lights were either on or dimmed.

The camera's field of view was set up to include the central ~ 1.5 m of the possible source locations. The source speed was up to $0.1\ \mathrm{ms}^{-1}$ (or ~ 0.7 pixels per frame in this camera configuration), which produced movement sequences that were slightly easier to track than the human sequences in Sections 4.1 and 4.3.2, which tended to have larger velocities and abrupt accelerations.

For each condition, ~ 40 s of training data was collected using three constant velocity passes of the AV source across the field of view of the camera. Then, 25 s of same condition test data was collected using a fixed pattern of movement (see Fig. 10), including fixed periods of (in)audibility and (in)visibility behind an occluding curtain. We tested performance using two approaches: 1) same condition testing, where the training data for the matching condition was used to train the model before testing on data from the same condition, and 2) cross-condition testing, where the training data for each visual appearance in the easiest condition (lights on, high volume, and high frequency) was used to train the model before testing on all the other conditions. These two evaluations quantify two aspects of performance: 1) the models' performance when trained appropriately under actual usage conditions (since a feature of our approach is rapid unsupervised learning of particular contexts) and 2) the models' performance when there is deviation between the training and usage scenarios.

4.2.2 Evaluation Results

We assume, for the purposes of evaluation, that the probabilistic tracker is required to make a single best guess of every quantity at every time and take the mode of the posterior distribution output at any point as its best answer. Fig. 10 details the distribution over the tracker outputs across the 24 test cases, with Figs. 10a and 10b, and 10c and 10d reporting the same-condition testing and cross-condition testing, respectively. The ground-truth position is illustrated by the plain black lines, and the ground-truth periods of video and audio occlusion are illustrated by the shaded bars below the plots. The distribution of outputs of the audio and video trackers is shown by the light/blue and medium/green shaded regions, respectively, and should be interpreted in the context of the occlusion periods. In Figs. 10a and 10c, the dark/red shaded region illustrates the output distribution for the IID pure-fusion model [6]. It almost entirely overlaps the (medium/green shaded) video region, as the video modality is dominant and deviates drastically from the ground truth (black line) during the entire video occlusion. In Figs. 10b and 10d, the dark/red shaded region illustrates the output for the filtered data association model developed in this paper. It relatively successfully follows the target using audio only during the initial part of the video occlusion but then fails once the audio occlusion begins. This is because based on its simple diffusion model of motion, it keeps predicting the same increasingly incorrect location, albeit with decreasing confidence. This continues until the video becomes available again, and tracking is regained.

The data illustrated in Fig. 10 is quantified in Table 2. For each model variant, we compute four performance measures:

1. *Track percentage*. This is the percentage of successfully tracked frames, defined as those for which the model output is within ± 10 pixels of the true target location.
2. *Accuracy*. This is the average absolute error in pixels between the model's estimate and the true location for those frames in which the target was tracked.

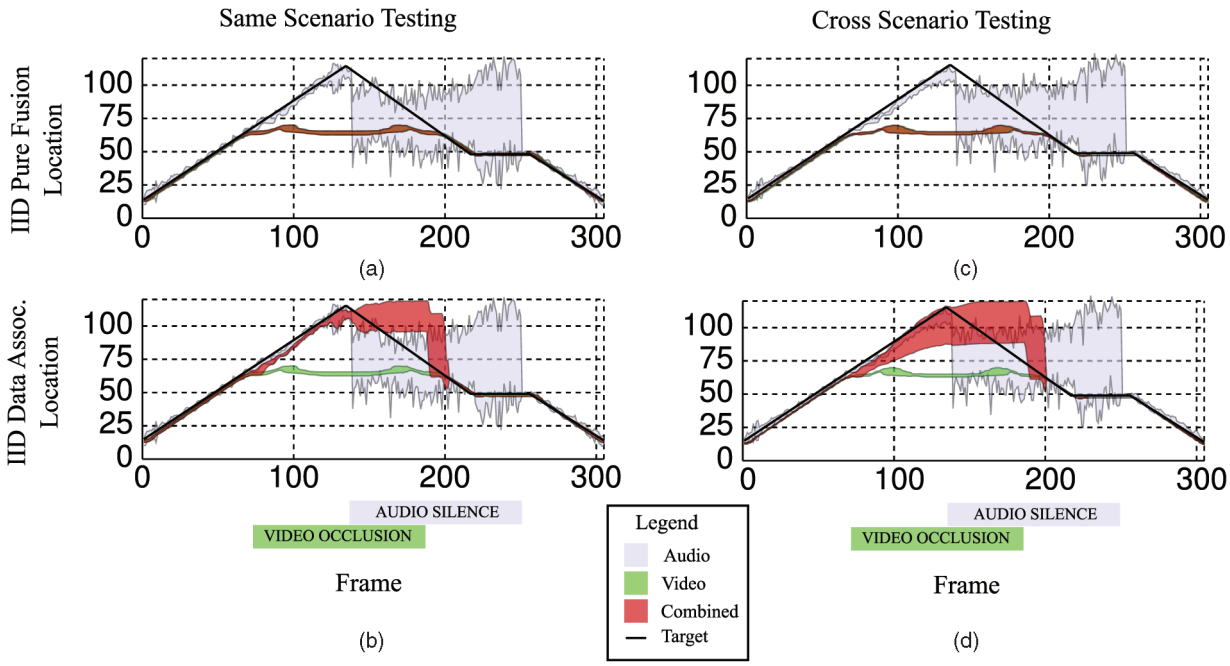


Fig. 10. Performance of AV tracking of the mechanically controlled target. Input is 24 test sequences of different statistics (see text). Ground truth is indicated by the plain black line. The light/blue and medium/green shaded regions indicate the distribution of tracking estimates over all recorded sequences for audio-only and video-only tracking, respectively. The dark/red shaded region indicates the distribution of estimates for (a) and (c) the IID pure-fusion model and (b) and (d) the filtered data association model. In (a) and (b), the models are tested on data of the same statistics for which they were trained. In (c) and (d), the base statistics are used to train the model, which is then tested on data of all the other statistics. Shaded bars under the plots indicate the regions of video and occlusion and audio silence in the sequence.

3. *Audio detection rate (ADR)*. This is the percentage of frames for which the audio was correctly identified as being audible or not.
4. *Video detection rate (VDR)*. This is the percentage of frames for which the video was correctly identified as being visible or not.

These are along the lines of standard evaluation measures for multimodal detection and tracking, for example, as formalized by the CLEAR evaluation campaign [1], [35].

A key aim of multimodal perception is to improve performance over any individual modality, so ideally, the combined models should perform better than the individual modalities. In this case, however, the pure-fusion models [6] do not outperform the unimodal tracking under any measure, because the fusion is dominated by the video, which can be unreliable during occlusion. In contrast, the models developed here, which try to infer the causal structure on the fly and generally succeed in doing so (see the ADR and VDR columns of Table 2) fuse the modalities only when appropriate and track most reliably. In particular, the filtered data association model that we develop here outperforms the IID pure-fusion model developed in [6] by some margin (see boldface numbers in Table 2). The performance reported in the left and right sections of Table 2 is for the same-condition and cross-condition testing, respectively. As expected, the same-condition performance is generally better than the cross-condition performance for each measure. However, it is worth noting that perfect cross-condition detection performance is not unambiguously positive, as eventually, we will want to discriminate among different sources of different statistics during multitarget tracking, as will be discussed in the Section 4.3.

4.2.3 Limitations of the Model

It is worth mentioning some limitations of the current model before continuing. So far, we have discussed only one-dimensional tracking in the horizontal plane. This is because the data that we are interested in often exhibits variability primarily in this plane and because the two-element microphone array only provides information in this plane, rendering multimodal cue combination only interesting in this plane. Using the techniques in Section 3.4.1 (see [13]), it is simple and efficient to compute visual likelihoods in both axes. However, this would render the tracking Markov model as presented here unfeasibly slow requiring, for example, sparse matrix techniques such as [21]. A stronger limitation is that the difficulty of representing visual rotation and scaling with TMG [13] precludes tracking these variations efficiently in our parametric framework. However, to some extent, the multimodal framework developed here can alleviate this problem, as tracking (in the horizontal plane at least) can continue based on the audio modality, even if vision fails due to excessive rotation or scaling.

Another class of potential problem relates to the unsupervised EM learning algorithm in the TMG framework [21], [13] rather than the tracking procedure. In trying to find a single set of parameters θ that maximize the likelihood of the data $\theta = argmax_\theta \prod^t p(X^t, Y^t|\theta)$, there may be many local maxima. For example, the foreground video model $(\mathrm{Z} = 1)$ can potentially learn a parameter μ to explain every video frame t as the stationary $(l^t = 0)$ (true) room background, with the (true) foreground user being explained away by noise Ψ on every frame. This could be more likely than the intended maxima if the following hold:

TABLE 2
Quantitative Evaluation of AV Tracking Results Using Mechanically Controlled Target

	SAME				CROSS			
	Track %	Accuracy	ADR %	VDR %	Track %	Accuracy	ADR %	VDR %
Aud Only	72.3 ± 2.5	2.57 ± 0.08	-	-	71.8 ± 3.1	2.01 ± 0.08	-	-
Vid Only	65.6 ± 1.1	2.52 ± 0.01	-	-	65.5 ± 0.6	3.10 ± 0.01	-	-
PF IID	$\mathbf{65.6} \pm 1.1$	2.52 ± 0.01	-	-	$\mathbf{65.5} \pm 0.6$	3.10 ± 0.01	-	-
PF Filt	65.6 ± 1.1	2.52 ± 0.01	-	-	65.5 ± 0.6	3.10 ± 0.01	-	-
DA IID	81.5 ± 2.4	2.67 ± 0.01	96.7 ± 10.2	100 ± 0	77.7 ± 6.5	3.19 ± 0.05	91.4 ± 20	100 ± 0
DA Filt	$\mathbf{86.3} \pm 2.6$	2.70 ± 0.01	96.7 ± 10.2	100 ± 0	$\mathbf{83.2} \pm 6.6$	3.24 ± 0.06	91.4 ± 20	100 ± 0

Results compare the percentage of frames with 1) successful tracking, 2) correct inference of audibility (ADR), and 3) visibility (VDR) of target. Only the last two methods computed ADR/VDR. For successfully tracked frames, the accuracy of tracking in terms of pixel error is also shown. Table SAME indicates tests performed using the input of the same statistics as the training data. Table CROSS indicates tests performed using one trained model and the input of all the other different statistics. The model in [6] corresponds to the row PF IID. See text for a detailed explanation of conditions.

1. The user's appearance area is very small compared to the background area.
2. The user is more frequently occluded than not in the training data.
3. The user is silent more frequently than not in the training data.
4. The actual background of the room is highly structured, making large translations difficult to explain by rotation, as required in TMG [21].

If many of these factors are true, inappropriate templates may be learned, and $\mathrm{Z}^t = 1$ may be inferred for all frames. Changing the parametric framework to one with a more explicit notion of layers [39], [20] may be necessary to entirely avoid these problems.

4.3 Inference for Multiple Sources

We have seen the benefits of a principled probabilistic approach to data association for user detection, robust tracking through occlusion, and multimodal user verification. However, the real value of explicit structural inference comes in multiobject scenarios, where the question of single-target user verification generalizes to the *who-said-what* problem. Exact inference unfortunately becomes exponentially more expensive in the maximum number of objects, as the objects' states become conditionally dependent, given their shared observations. Nevertheless, we shall see that with some small changes to the model as described in (8) and Fig. 6, we can efficiently approximate inference in the multitarget scenario and solve the who-said-what problem.

4.3.1 Multitarget Tracking Framework

The first approximation to multitarget inference is simply to ignore the conditional dependency between the latent states of each object. Two separate instances of the model (such as that of (8) and Fig. 6) can each be initially trained with data containing a target of interest. In other words, data D_A containing samples of target A is used to train a Model A using EM with ML parameters $\theta_A = \mathrm{argmax}_{\theta_A} p(D_A|\theta_A)$, and Model B learns the ML parameters θ_B from data D_B containing samples of target B, $\theta_B = \mathrm{argmax}_{\theta_B} p(D_B|\theta_B)$. Once trained, these models can perform multitarget tracking and scene understanding by a simultaneous but independent inference such that Model A computes $p(l_A^t, \mathrm{W}_A^t, \mathrm{Z}_A^t | D^{1:t}, \theta_A)$ and Model B computes $p(l_B^t, \mathrm{W}_B^t, \mathrm{Z}_B^t | D^{1:t}, \theta_B)$. This is a linear, rather than exponential, cost in the number of targets.

The suitability of this approach depends on to what extent data from each target behaves like explainable noise from the perspective of the tracker concerned with the other target. This assumption does not quite hold, given the model as introduced in Section 3 and trained as described in Section 4. The main reasoning behind this is the fact that after learning, the parameters in θ_i describe two classes of audio data: the "foreground" speech of large amplitude and associated source location and the "background" office noise of smaller amplitude and uncorrelated source location. In the multitarget scenario, there are now three empirical classes of audio data: the foreground associated speech (generated from the target of interest), the foreground disassociated speech (generated from another target not of interest, hence with τ uncorrelated with l), and the background office noise.

To decide if we associate a given frame of audio data $(\mathbf{x}_1, \mathbf{x}_2)^t$ with its target, Model A computes

$$p(\mathrm{W}_A^t | D^{1:t}, \theta_A) = \sum_{\mathrm{Z}_A, l_A} p(l_A^t, \mathrm{W}_A^t, \mathrm{Z}_A^t | D^{1:t}, \theta_A).$$

This depends on two important factors in the generative model. First is the three-way *match* between the peak of this likelihood as a function of l, the prior predicted location probability $p(l^t | D^{1:t-1})$, and the likelihood of the video observation $p(\mathbf{y}^t | l_A, \mathrm{Z}_A, \theta_A)$ (this is exactly the point that was introduced in Section 2 and illustrated in Figs. 2 and 3). Second, the association depends on the template match as specified by the likelihood of the audio data under the background and foreground distributions unique to the AV model:

$$p(\mathbf{x}_1^t, \mathbf{x}_2^t | l_A, \mathrm{W}_A, \theta_A) = \int_{\mathbf{a}} \sum_{\tau} \mathcal{N}(\mathbf{x}_1 | \mathbf{a}, \upsilon_1)^w \mathcal{N}(\mathbf{x}_1 | \mathbf{0}, \sigma_1)^{\overline{w}} \cdot \mathcal{N}(\mathbf{x}_2 | T_\tau \mathbf{a}, \upsilon_2)^w \mathcal{N}(\mathbf{x}_2 | \mathbf{0}, \sigma_2)^{\overline{w}} \mathcal{N}(\tau | \alpha l + \beta, \omega). \quad (31)$$

The new empirical class of data, that is, disassociated speech, will be probable under the audio template likelihood model. At the same time, it will be unlikely in terms of the match between the *shape* of this likelihood: that of the

video and predictive distribution from the Markov chain. In practice, this means that the disassociated speech would frequently be inappropriately classified as associated speech.[3] Therefore, we introduce the second background model to account properly for all three classes of audio data that are now present. Conveniently, the additional model only needs parameters already determined during learning. Let W now be 3D multinomial, defining the following audiomodality likelihoods:

$$\begin{aligned} p(\mathbf{x}_1, \mathbf{x}_2|\mathbf{a}, \tau, \mathrm{W}=1) &= \mathcal{N}(\mathbf{x}_1|\mathbf{a}, \upsilon_1)\mathcal{N}(\mathbf{x}_2|T_\tau \mathbf{a}, \upsilon_2), \\ p(\tau|l, \mathrm{W}=1) &= \mathcal{N}(\tau|\alpha l + \beta, \omega), \end{aligned} \tag{32}$$

$$\begin{aligned} p(\mathbf{x}_1, \mathbf{x}_2|\mathbf{a}, \tau, \mathrm{W}=2) &= \mathcal{N}(\mathbf{x}_1|\mathbf{a}, \upsilon_1)\mathcal{N}(\mathbf{x}_2|T_\tau \mathbf{a}, \upsilon_2), \\ p(\tau|\mathrm{W}=3) &= \mathcal{U}(\tau), \end{aligned} \tag{33}$$

$$\begin{aligned} p(\mathbf{x}_1, \mathbf{x}_2|\mathbf{a}, \tau, \mathrm{W}=3) &= \mathcal{N}(\mathbf{x}_1|\mathbf{0}, \sigma_1)\mathcal{N}(\mathbf{x}_2|\mathbf{0}, \sigma_2), \\ p(\tau|\mathrm{W}=3) &= \mathcal{U}(\tau). \end{aligned} \tag{34}$$

The foreground model $\mathrm{W}=1$ is unchanged (32), the first background model $\mathrm{W}=2$ (33) now accounts for signals with the statistics of speech but without any expected correlation with the predicted location or the likelihood of the video, and the second background model $\mathrm{W}=3$ (33) is also unchanged from before, accounting for the background office noise.

4.3.2 Multitarget Tracking: Detailed Example

The results for such a multitarget scenario are illustrated in Fig. 11. In this scene, two users are having a discussion while moving around, occasionally passing in front of each other. The raw input waveform and video data are illustrated in Figs. 11a and 11b. Models A and B have previously been trained independently on data (similar to that of Figs. 7a and 7b) containing their respective users and learned, among other parameters $\theta_{A,B}$, the video templates $\mu_{A,B}$ shown in Figs. 11c and 11d. The trained models each now report the posterior distribution over location and data association for their user u, $p(\mathrm{W}_u^t, \mathrm{Z}_u^t, l_u^t|D^{1:t}, \theta_u)$.

The smoothed posterior distribution over audio association $p(\mathrm{W}_u=1|D)$ is shown in Fig. 11g, with a gray/green line for user 1 and a black line for user 2. The turn-taking behavior in the conversation is clear with the alternating modes in the distribution for each. The posterior over video association $p(\mathrm{Z}_u|D)$ is shown in Fig. 11h. The initial presence of user 1 in the video is indicated by the initially high value for $p(\mathrm{Z}_1|D)$, and the subsequent entrance of user 2 is indicated by the rising initial value for $p(\mathrm{Z}_2|D)$. The fact that the subsequent occlusions as the users pass each other in the scene are correctly inferred is clear by the later dips in the line. Finally, the MAP location of each user is illustrated in Fig. 11i, along with the audio and video likelihood modes for each model. Similar to the situation in Section 4.1, during visual occlusion, the video likelihood modes are quite spurious, but the detection and tracking functionality ensures that the spurious modes are ignored until the user is visible again.

3. The observation likelihood under the background model $\overline{w}$ is very low, as it is implausible that every component of the two 1,000-dimensional background Gaussians $\mathcal{N}(\mathbf{x}_1|\mathbf{0}, \sigma_1)\mathcal{N}(\mathbf{x}_2|\mathbf{0}, \sigma_2)$ simultaneously become large. This is a much stronger effect than the mismatch in shape between the foreground likelihood and the video and predictive distributions, which occur only in one dimension.

An important and novel feature of this framework is that segmentation of the original raw speech data $\mathbf{x}_1$ and $\mathbf{x}_2$ is now provided, that is, as a byproduct of inference, by the posterior probability of audio association $p(\mathrm{W}_u|D)$. That is, $p(\mathrm{W}_u^t=1|D^{1:t})$ defines the posterior probability that the speech at time t : $(\mathbf{x}_1^t, \mathbf{x}_2^t)$ originated in user u. It is therefore the probabilistic answer to the question of who uttered the current frame of speech. The speech segments uttered by each user are extracted from the raw data using $p(\mathrm{W}_u^t=1|D^{1:t})$, as illustrated in Figs. 11e and 11f. This is the solution to the *who-said-what* problem.[4] In contexts such as conversation understanding, transcription, and summarization [17], the segmented speech signals could then be passed on to a speech processing system to produce a speaker-labeled transcription.

4.3.3 Multitarget Tracking: Quantitative Evaluation

In this section, we summarize the quantitative performance of the models in a multitarget tracking context. We recorded five multiparty conversation video sequences of approximately 1 minute each along the lines of the one examined in detail in Section 4.3.2. The sequences included some different room configurations and users: this necessitated the learning of the different AV appearances. To create the ground truth, we manually labeled the location, visibility, and speaking status of the users in each frame. Given the ground-truth data, we were able to quantify performance by using a similar procedure to that described in Section 4.2.2.

Table 3 details the tracking performance of the models in this multitarget scenario averaged over all the recorded sequences. Based on the key measure of the percentage of successfully tracked frames (Track %), the audio-only tracking performance is much lower than that in Table 2. This is because the speech signal is less precisely localizable and more intermittent than the noise signal used in Section 4.2.2. Nevertheless, combining the audio and video modalities with structure inference allows the filtered data association model to perform better than either modality alone, as well as better than the pure-fusion model [6].

Next, we evaluate the AV association performance. The earlier model variants do not compute this, so we focus on the performance of the final filtered data association model. The audio model now has three possible structures W. The total error can first be computed as the percentage of frames for which the ground truth W^{gt} and the model's MAP estimate W^{est} do not match $\mathrm{W}_u^{gt} \neq \mathrm{W}_u^{est}$. Since we are mostly interested in detecting the correct speaker $\mathrm{W}_u=1$ and not the nature of the negatives ($\mathrm{W}_u=2$ versus $\mathrm{W}_u=3$), we combine the two negative categories when computing the effective error rate. The effective error rate can then be further broken down into false positives, that is, reporting that user u is speaking when he is actually silent ($\mathrm{W}_u=1$, but $\mathrm{W}_u^{gt}=2,3$), and false negatives, that is, reporting that user u is silent when he is actually speaking ($\mathrm{W}_u=2,3$, but $\mathrm{W}_u^{gt}=1$). The detection rates are computed similarly for the video modality. The results are reported in Table 4.

In this multiparty conversation context, an interesting quantity is the accuracy with which the model can assign speech segments to the users. Therefore, in Table 5, we also report the average confusion matrix between the actual and

4. See video sequences on the project site http://www.ipab.inf.ed.ac.uk/slmc/projects/CueIntegrate.html.

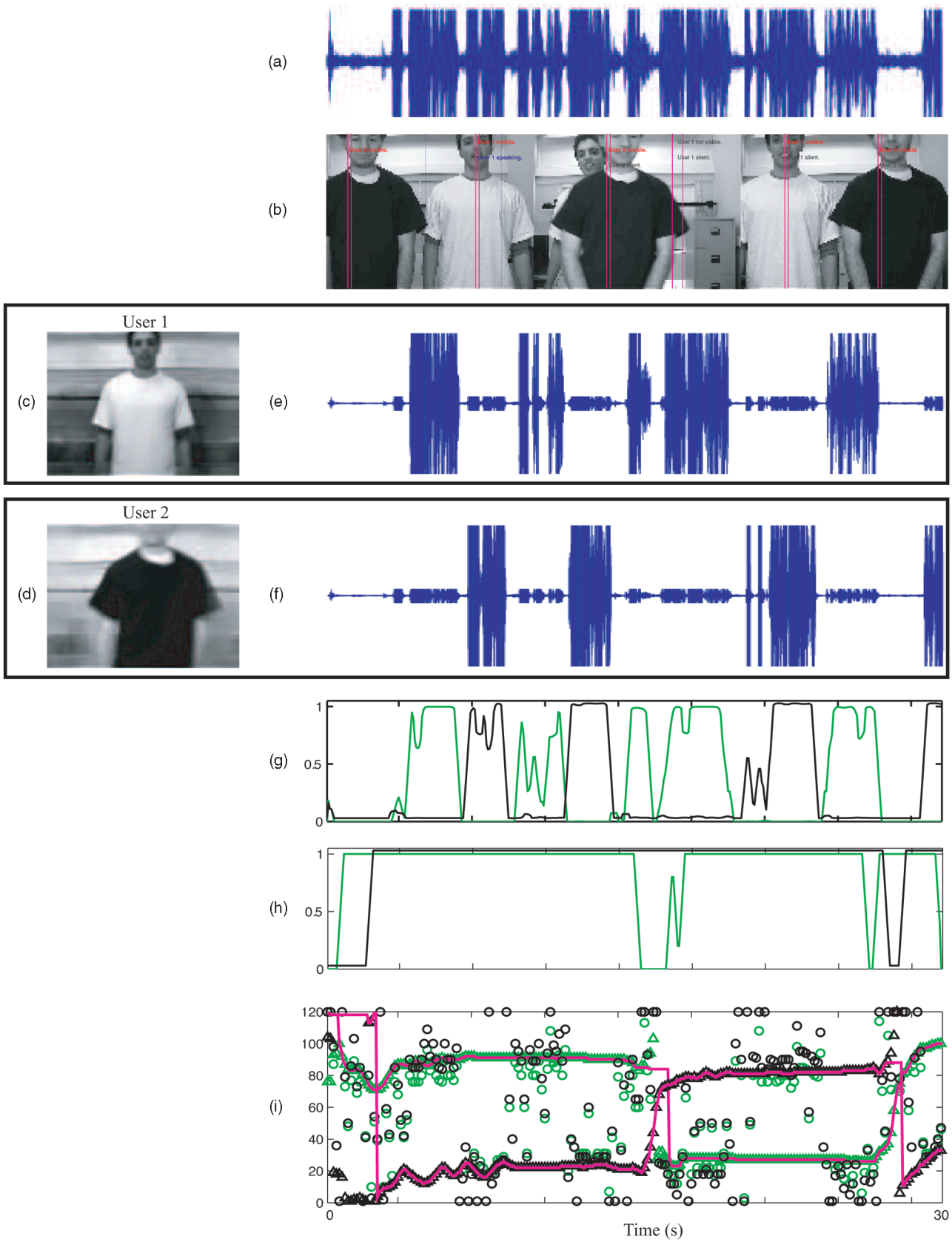


Fig. 11. AV multiobject tracking and scene understanding results. (a) Raw audio data and (b) sample video frames from a sequence where two users are conversing and moving around, occasionally occluding each other. (c) and (d) Learned templates for the two users. (e) and (f) Speech segments inferred to belong to each user. Posterior probability of (g) audibility and (h) visibility for users 1 (gray/green) and 2 (black). (i) Multiuser tracking. Audio likelihood peaks are shown as circles, and video likelihood peaks are shown as triangles. MAP locations are shown by the two dark/purple lines.

TABLE 3
Summary of Multiuser Tracking Performance

Model	Track %	Accuracy (Pixels)
AO	28.9 ± 7.0	4.33 ± 0.52
VO	86.3 ± 19.1	1.99 ± 1.06
PF IID	$\mathbf{86.3} \pm 19.1$	1.99 ± 1.06
PF Filt	86.4 ± 19.1	1.99 ± 1.06
DA IID	86.2 ± 15.2	2.01 ± 1.07
DA Filt	$\mathbf{88.7} \pm 12.6$	2.04 ± 1.14

Track percent indicates the percentage of time that the tracker's output was on target, that is, within ± 10 pixels of the true target location. Accuracy indicates the absolute error in pixels of the tracker for the correctly tracked frames.

TABLE 4
User Detection Rate in a Multiparty Scenario

	ADR	VDR
Total Error	$19.0 \pm 8.5\%$	-
Effective Error	$12.9 \pm 7.5\%$	3.0 ± 4.5 %
False Pos	$4.7 \pm 6.2\%$	$1.3 \pm 4.0\%$
False Neg	$8.1 \pm 2.7\%$	$1.7 \pm 2.9\%$

TABLE 5
Confusion Matrix for Multiuser Speech Segmentation

		Actual		
		U1	U2	None
Reported	U1	**72.6**$\pm$16.0%	$9.5 \pm 15.0\%$	5.5$\pm$5.3%
	U2	$11.0 \pm 13.4\%$	**74.6**$\pm$21.7%	3.3$\pm$3.3%
	None	$16.4 \pm 8.6\%$	$15.9 \pm 9.3\%$	**91.2**$\pm$7.6%

reported speakers of each segment in terms of containing speech from user 1, user 2, or neither. The model performs well, correctly assigning at least 72 percent of the speech segments to the user uttering them.

4.4 Summary

In this section, we have illustrated the application of the ideas introduced in Sections 2 and 3 to a real AV scene understanding problem. Multisensory detection, verification, and robust tracking through occlusion of either or both modalities are achieved through the inference of latent state and structure. The inference turns out to depend on a combination of three effects: the correlation between the shape of the observation likelihoods in each modality, the correlation between the shape of the observation likelihoods and the predictive distribution, and the goodness of the template match in each modality.

The multitarget data association problem is more interesting, as the solution to it represents explicit relational knowledge of who was present (visible) when and who said what when. While expensive to compute exactly, in this application, an independence approximation in which the background models for each user explain data generated by the other user turns out to be sufficient for robust multitarget tracking and data association. A probabilistic segmentation of the speech is achieved as a byproduct of the explicit computation of data association.

5 Discussion

In this paper, we introduced a principled formulation of multisensory perception and tracking in the framework of Bayesian inference and model selection in probabilistic graphical models. Pure-fusion multisensor models have previously been applied in machine perception applications and in 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. Moreover, in many cases, inferring data 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. More significantly, to understand a multiparty conversation, the speech segments need to be correctly associated with person identity. In our application, the model computes which observations arise from which sources by explicitly inferring association, so it can, for example, start a recording when the user enters the scene or begins speaking and segment the speech in a multiparty conversation.

5.1.1 Related Research

While we have discussed relevant previous research in Section 1, it is worth contrasting our study against some related pieces of recent and ongoing work. In radar tracking and association, some work [36] uses similar techniques to ours; however, popular methods [3] tend to be more heuristic, necessarily use stronger assumptions and approximations (for example, Gaussian posteriors), and use highly preprocessed point-input data. One interesting contrast between these candidate-detection-based approaches and our generative model approach is that we avoid the expensive within-modality data association problem typical of radar. This also enables the use of signature or template information in a unified way, along with cross-modality correlation during inference, which is exploited to a good effect in our AV application.

In AV processing, Siracusa and Fisher III [33] independently propose a model that computes association between two speakers and their speech segments by inferring the presence or absence of conditional dependencies. However, this model is specific to this task and does not handle the full tracking and AV template learning problem that we address simultaneously here. Another interesting model is [14], which uses particle filter inference on AV data to perform tracking and AV speech association. The less constrained particle filter framework used in [14] allows handling of birth and death processes for many sources, a topic that we have not addressed. However, vision and audition play unequal roles in [14], with sources being "eliminated" if not visible. This precludes tracking through occlusion, as we illustrate in this paper. Moreover, [14] does not address learning and therefore requires hand calibration of the individual modality trackers and AV connection.

In computer vision, [39] and [20] describe techniques related to ours for the unsupervised learning and tracking of multiple objects in video by using greedy and variational inference approximations, respectively. These do not require the independent learning for each target used in our framework. However, in using only one modality, [39] and [20] avoid the multimodal data association problem that we address here.

5.1.2 Future Work

Investigations of human multisensory perception have reported robustness to discrepant cues [10] but principled theory to explain that this has been lacking. We envisage that our theory can be used to understand a much wider

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 that we are currently investigating. Indeed, very recent research has suggested that this is indeed the case for AV perception in humans [22].

In the context of machine perception, the framework described generalizes existing pure-fusion models and using a single probabilistic framework, 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 the source state and multitarget data association for real problems is potentially even real time. Our future research theme is to integrate our existing work on sensorimotor control [38] with these probabilistic perceptual models to extend them into the domain of active perception.

APPENDIX A
MODEL UPDATE EQUATIONS

In this section, we list the derived updates for all model parameters as required for the EM algorithm.

A.1 Video Appearance Model Updates

Here, we make use of the sufficient statistics from the video inference $\mu^t_{\mathbf{v}|\mathbf{y},l,z}$ and $\nu_{\mathbf{v}}|z$, as computed in (10) and (11). Define, for convenience, $N_z = \sum_t p(z^t|D^{1:T})$ and $N_{\overline{z}} = \sum_t p(\overline{z}^t|D^{1:T})$ to be the total weight of the associated and disassociated video frames, respectively, in the training sequence:

$$\mu \leftarrow \frac{1}{N_z}\sum_{t,l} p(l^t, z^t|D^{1:T})\mu^t_{\mathbf{v}|\mathbf{y},l,z},$$

$$\phi^{-1} \leftarrow \frac{1}{N_z}\sum_{t,l} p(l^t, z^t|D^{1:T}) \cdot \mathrm{Diag}\Big((\mu^t_{\mathbf{v}|\mathbf{y},l,z} - \mu)(\mu^t_{\mathbf{v}|\mathbf{y},l,z} - \mu)^T + (\nu_{\mathbf{v}|z})^{-1}\Big),$$

$$\Psi^{-1} \leftarrow \frac{1}{N_z}\sum_{t,l} p(l^t, z^t|D^{1:T}) \cdot \mathrm{Tr}\Big((\mathbf{y}^t - \mathbf{T}_l\mu^t_{\mathbf{v}|\mathbf{y},l,z})(\mathbf{y}^t - \mathbf{T}_l\mu^t_{\mathbf{v}|\mathbf{y},l,z})^T + (\nu_{\mathbf{v}|z})^{-1}\Big),$$

$$\gamma \leftarrow \frac{1}{N_z N_{\mathbf{y}}}\sum_t p(\overline{z}^t|D^{1:T})\sum_i \mathbf{y}^t_{[i]},$$

$$\epsilon^{-1} \leftarrow \frac{1}{N_z N_{\mathbf{y}}}\sum_t p(\overline{z}^t|D^{1:T})(\mathbf{y}^t - \gamma)^2.$$

Here, $N_{\mathbf{y}}$ is the total number of pixels per frame, and the inner product $\mathbf{x}^T\mathbf{x}$ is written as $\mathbf{x}^2$.

A.2 Audio Appearance Model Updates

Here, we make use of the sufficient statistics from the audio inference $\mu_{\mathbf{a}|\mathbf{x},\tau,w}$ and $\nu_{\mathbf{a}}$, as computed in (12) and (13). The full posterior over the interaural time delay, as well as location and association

$$p(\tau^t, l^t, \mathrm{W}^t, \mathrm{Z}^t|D^{1:T}) = p(\tau^t|l^t, \mathrm{W}^t, \mathrm{D}^{1:T})p(l^t, \mathrm{W}^t, \mathrm{Z}^t|D^{1:T}),$$

as computed by (14) and (24), is also used. Define, for convenience, $N_w = \sum_t p(w^t|D^{1:T})$ and $N_{\overline{w}} = \sum_t p(\overline{w}^t|D^{1:T})$ to be the total weight of the associated and disassociated video frames, respectively, in the training sequence. $N_{\mathbf{x}}$ is the total number of audio samples per frame:

$$\lambda_1 \leftarrow \frac{\sum_{t,\tau} p(\tau^t, w^t|D^{1:T})\mathbf{x}_1^T\mu_{\mathbf{a}|\mathbf{x},\tau,w}}{\sum_{t,\tau} p(\tau^t, w^t|D^{1:T})(\mu^t_{\mathbf{a}|\mathbf{x},\tau,w})^2 + N_w\mathrm{Tr}(\nu^{-1}_{\mathbf{a}|w})},$$

$$\lambda_2 \leftarrow \frac{\sum_{t,\tau} p(\tau^t, w^t|D^{1:T})\mathbf{x}_2^T\mathbf{T}_\tau\mu_{\mathbf{a}|\mathbf{x},\tau,w}}{\sum_{t,\tau} p(\tau^t, w^t|D^{1:T})(\mu^t_{\mathbf{a}|\mathbf{x},\tau,w})^2 + N_w\mathrm{Tr}(\nu^{-1}_{\mathbf{a}|w})},$$

$$\upsilon_1^{-1} \leftarrow \frac{1}{N_w N_{\mathbf{x}}}\Bigg(\sum_{t,\tau} p(\tau^t, w^t|D^{1:T}) \cdot (\mathbf{x}_1^t - \lambda_1\mu^t_{\mathbf{a}|\mathbf{x},\tau,w})^2\Bigg) + N_w\lambda_1^2\mathrm{Tr}(\nu^{-1}_{\mathbf{a}|w}),$$

$$\upsilon_2^{-1} \leftarrow \frac{1}{N_w N_{\mathbf{x}}}\Bigg(\sum_{t,\tau} p(\tau^t, w^t|D^{1:T}) \cdot (\mathbf{x}_2^t - \lambda_2\mathbf{T}_\tau\mu^t_{\mathbf{a}|\mathbf{x},\tau,w})^2\Bigg) + N_w\lambda_2^2\mathrm{Tr}(\nu^{-1}_{\mathbf{a}|w}),$$

$$\eta^{-1} \leftarrow \frac{1}{N_w N_{\mathbf{x}}}\Bigg(\sum_{t,\tau} p(\tau^t, w^t|D^{1:T})(\mu^t_{\mathbf{a}|\mathbf{x},\tau,w})^2 + \mathrm{Tr}(\nu^{-1}_{\mathbf{a}|w})\Bigg),$$

$$\sigma_1^{-1} \leftarrow \frac{1}{N_{\overline{w}} N_{\mathbf{x}}}\sum_t p(\overline{w}^t|D^{1:T})(\mathbf{x}_1^t)^2,$$

$$\sigma_2^{-1} \leftarrow \frac{1}{N_{\overline{w}} N_{\mathbf{x}}}\sum_t p(\overline{w}^t|D^{1:T})(\mathbf{x}_2^t)^2.$$

A.3 Multimodal Updates

Suppressing indexing by time t for clarity, the parameters of the AV link are updated as follows:

$$\beta \leftarrow \frac{1}{N_w}\Bigg(\sum_{t,\tau,l} Q\cdot\tau - \alpha\sum_{t,\tau,l} Q\cdot l\Bigg),$$

$$\alpha \leftarrow \frac{\sum_{t,\tau,l} Q\Big(l\tau - l\frac{1}{N_w}\sum_{t,\tau,l} Q\cdot\tau\Big)}{\sum_{t,\tau,l} Q\cdot l^2 - \Big(\sum_{t,\tau,l} Q\cdot l\cdot\Big(\frac{1}{N_w}\sum_{t,\tau,l} Q\cdot l\Big)\Big)},$$

$$\omega^{-1} \leftarrow \frac{1}{N_w}\sum_{t,\tau,l} Q\big(\tau^2 - 2\tau\alpha l - 2\tau\beta + \alpha^2 l^2 + 2\alpha l\beta + \beta^2\big),$$

where $Q \triangleq p(\tau^t, l^t, w^t|D^{1:T})$.

A.4 Markov Chain Updates

To compute the updates for the Markov chain parameters, we make use of the sufficient statics α (5) and γ (6) from the inference as follows:

$$\mathbf{\Gamma}_{[i,j]} \leftarrow \frac{\sum_t p(l^t, l^{t+1}|D^{1:T})_{[i,j]}}{\sum_t \gamma(l^t)_{[i]}},$$

$$\Theta_{[i,j]} \leftarrow \frac{\sum_t p(W^t, W^{t+1}|D^{1:T})_{[i,j]}}{\sum_t \gamma(W^t)_{[i]}},$$

$$\Omega_{[i,j]} \leftarrow \frac{\sum_t p(Z^t, Z^{t+1}|D^{1:T})_{[i,j]}}{\sum_t \gamma(Z^t)_{[i]}},$$

where $p(l^t, l^{t+1}|D^{1:T}) = \frac{\alpha(l^t)p(D^{t+1}|l^{t+1})\gamma(l^t)\Gamma_{[l^t,l^{t+1}]}}{\alpha(l^{t+1})}$.

REFERENCES

[1] CLEAR 2006 Evaluation and Workshop Campaign, http://www.clear-evaluation.org/, Apr. 2006.

[2] D. Alais and D. Burr, "The Ventriloquist Effect Results from Near-Optimal Bimodal Integration," *Current Biology,* vol. 14, no. 3, pp. 257-262, Feb. 2004.

[3] 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,* vol. 20, no. 8, pp. 37-56, 2005.

[4] Y. Bar-Shalom and E. Tse, "Tracking in a Cluttered Environment with Probabilistic Data Association," *Automatica,* vol. 11, pp. 451-460, 1975.

[5] P.W. Battaglia, R.A. Jacobs, and R.N. Aslin, "Bayesian Integration of Visual and Auditory Signals for Spatial Localization," *J. Optical Soc. Am. A: Optics, Image Science, and Vision,* vol. 20, no. 7, pp. 1391-1397, July 2003.

[6] M.J. Beal, N. Jojic, and H. Attias, "A Graphical Model for Audiovisual Object Tracking," *IEEE Trans. Pattern Analysis and Machine Intelligence,* vol. 25, no. 7, pp. 828-836, July 2003.

[7] J. Bilmes, "Dynamic Bayesian Multinets," *Proc. 16th Ann. Conf. Uncertainty in Artificial Intelligence (UAI),* 2000.

[8] C. Boutilier, N. Friedman, M. Goldszmidt, and D. Koller, "Context-Specific Independence in Bayesian Networks," *Proc. 12th Ann. Conf. Uncertainty in Artificial Intelligence (UAI),* 1996.

[9] Y. Chen and Y. Rui, "Real-Time Speaker Tracking Using Particle Filter Sensor Fusion," *Proc. IEEE,* vol. 92, no. 3, pp. 485-494, Mar. 2004.

[10] M.O. Ernst and M.S. Banks, "Humans Integrate Visual and Haptic Information in a Statistically Optimal Fashion," *Nature,* vol. 415, pp. 429-433, 2002.

[11] J.W. Fisher III and T. Darrell, "Speaker Association with Signal-Level Audiovisual Fusion," *IEEE Trans. Multimedia,* vol. 6, no. 3, pp. 406-413, 2004.

[12] T.E Fortmann, Y. Bar-Shalom, and M. Scheffe, "Sonar Tracking of Multiple Targets Using Joint Probabilistic Data Association," *IEEE J. Oceanic Eng.,* vol. 8, pp. 173-184, 1983.

[13] B. Frey and N. Jojic, "Transformation-Invariant Clustering Using the EM Algorithm," *IEEE Trans. Pattern Analysis and Machine Intelligence,* vol. 25, no. 1, pp. 1-17, Jan. 2003.

[14] D. Gatica-Perez, G. Lathoud, J.-M. Odobez, and I.A. McCowan, "Audio-Visual Probabilistic Tracking of Multiple Speakers in Meetings," *IEEE Trans. Audio, Speech, and Language Processing,* vol. 15, pp. 601-616, 2007.

[15] D. Geiger and D. Heckerman, "Knowledge Representation and Inference in Similarity Networks and Bayesian Multinets," *Artificial Intelligence,* vol. 82, pp. 45-74, 1996.

[16] Z. Ghahramani and M. Jordan, "Factorial Hidden Markov Models," *Machine Learning,* vol. 29, pp. 245-273, 1997.

[17] T. Hain, J. Dines, G. Garau, M. Karafiat, D. Moore, V. Wan, R. Ordelman, and S. Renals, "Transcription of Conference Room Meetings: An Investigation," *Proc. Ninth European Conf. Speech Comm. and Technology,* 2005.

[18] J. Hershey and J.R. Movellan, "Using Audio-Visual Synchrony to Locate Sounds," *Advances in Neural Information Processing Systems,* 1999.

[19] R.A. Jacobs, "Optimal Integration of Texture and Motion Cues to Depth," *Vision Research,* vol. 39, no. 21, pp. 3621-3629, Oct. 1999.

[20] N. Jojic and B. Frey, "Learning Flexible Sprites in Video Layers," *Proc. IEEE Conf. Computer Vision and Pattern Recognition (CVPR '01),* vol. 1, 2001.

[21] N. Jojic, N. Petrovic, B.J. Frey, and T.S. Huang, "Transformed Hidden Markov Models: Estimating Mixture Models of Images and Inferring Spatial Transformations in Video Sequences," *Proc. IEEE Conf. Computer Vision and Pattern Recognition (CVPR '00),* vol. 2, pp. 26-33, June 2000.

[22] K.P. Kording, U. Beierholm, W.J. Ma, S. Quartz, J.B Tenenbaum, and L. Shams, "Causal Inference in Multisensory Perception," *PLoS ONE,* vol. 2, no. 9, p. 943, 2007.

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

[24] V.K. Mansinghka, C. Kemp, J.B. Tenenbaum, and T.L. Griffiths, "Structured Priors for Structure Learning," *Proc. 22nd Conf. Uncertainty in Artificial Intelligence (UAI),* 2006.

[25] A.V. Nefian, L. Liang, X. Pi, X. Liu, and K. Murphy, "Dynamic Bayesian Networks for Audio-Visual Speech Recognition," *EURASIP J. Applied Signal Processing,* vol. 11, pp. 1-15, 2002.

[26] P. Perez, J. Vermaak, and A. Blake, "Data Fusion for Visual Tracking with Particles," *Proc. IEEE,* vol. 92, no. 3, pp. 495-513, 2004.

[27] C. Rasmussen and G.D. Hager, "Probabilistic Data Association Methods for Tracking Complex Visual Objects," *IEEE Trans. Pattern Analysis and Machine Intelligence,* vol. 23, pp. 560-576, 2001.

[28] G.H. Recanzone, "Auditory Influences on Visual Temporal Rate Perception," *J. Neurophysiology,* vol. 89, pp. 1078-1093, 2003.

[29] D. Serby, E.-K. Meier, and L. Van Gool, "Probabilistic Object Tracking Using Multiple Features," *Proc. 17th Int'l Conf. Pattern Recognition (ICPR),* 2004.

[30] L. Shams, Y. Kamitani, and S. Shimojo, "Illusions: What You See Is What You Hear," *Nature,* vol. 408, p. 788, Dec. 2000.

[31] L. Shams, W.J. Ma, and U. Beierholm, "Sound-Induced Flash Illusion as an Optimal Percept," *Neuroreport,* vol. 16, no. 17, pp. 1923-1927, 2005.

[32] R. Silva and R. Scheines, "Bayesian Learning of Measurement and Structural Models," *Proc. 23rd Int'l Conf. Machine Learning (ICML),* 2006.

[33] M.R. Siracusa and J.W. Fisher III, "Dynamic Dependency Tests: Analysis and Applications to Multi-Modal Data Association," *Proc. 11th Int'l Conf. Artificial Intelligence and Statistics (AIStats),* 2007.

[34] M. Slaney and M. Covell, "Facesync: A Linear Operator for Measuring Synchronization of Video Facial Images and Audio Tracks," *Advances in Neural Information Processing Systems,* 2000.

[35] *Multimodal Technologies for Perception of Humans,* LNCS 4122, R. Stiefelhagen and J. Garofolo, eds., Springer, 2007.

[36] L.D. Stone, C.A. Barlow, and T.L. Corwin, *Bayesian Multiple Target Tracking.* Artech House, 1999.

[37] J. Vermaak, S.J. Godsill, and P. Perez, "Monte Carlo Filtering for Multi Target Tracking and Data Association," *IEEE Trans. Aerospace and Electronic Systems,* vol. 41, no. 1, pp. 309-332, Jan. 2005.

[38] S. Vijayakumar, J. Conradt, T. Shibata, and S. Schaal, "Overt Visual Attention for a Humanoid Robot," *Proc. IEEE/RSJ Int'l Conf. Intelligence in Robotics and Systems (IROS),* 2001.

[39] C.K.I. Williams and M.K Titsias, "Greedy Learning of Multiple Objects in Images Using Robust Statistics and Factorial Learning," *Neural Computation,* vol. 16, no. 5, pp. 1039-1062, May 2004.

Timothy Hospedales received the BA degree in computer science from the University of Cambridge in 2001 and the MSc degrees in informatics and neuroinformatics from the University of Edinburgh in 2002 and 2003, respectively. He is currently working toward the PhD degree on Bayesian models of sensor combination in the Neuroinformatics and Computational Neuroscience Doctoral Training Center (Neuroinformatics DTC) and the Institute of Perception, Action, and Behavior, School of Informatics, University of Edinburgh, United Kingdom. His research interests include Bayesian inference, machine learning, robotics, and theoretical neuroscience.

Sethu Vijayakumar received the PhD degree in computer science and engineering from Tokyo Institute of Technology in 1998. He is the director of the Institute of Perception, Action, and Behavior, School of Informatics, University of Edinburgh, United Kingdom. Since 2007, he has held a fellowship with the Royal Academy of Engineering in Learning Robotics, cosponsored by Microsoft Research, Cambridge. He also holds additional appointments as an adjunct faculty at the University of Southern California, Los Angeles, a research scientist in the ATR Computational Neuroscience Laboratories, Kyoto, Japan, and a visiting research scientist at the RIKEN Brain Science Institute, Tokyo. His research interests include a broad interdisciplinary curriculum ranging from statistical machine learning, robotics, planning, and optimization in autonomous systems to motor control and computational neuroscience.