Bayesian Multisensory Perception

Timothy Hospedales, University of Edinburgh

October 31, 2007

Introduction

An important aspect of a cognitive system’s perceptual function is its ability to intelligently combine many disparate sources of information. For example, suppose you are trying to locate your dog who has run away in the park. During the day you might visually search the horizon for movement, while during the night you might follow its bark, or on a foggy evening you might do a combination of both. This example is a problem of inferring a continuous state (the pet’s location) on the basis of two observations (vision and audition) which may be independently degraded (e.g. by darkness/fog and sound dispersion respectively).

Another aspect to this problem is the relation between the cues, which is indirectly related to the state of interest. Suppose while searching: i) you see motion on the horizon, but of a different color than your dog’s, ii) you hear a bark, but in a different tone to your dog’s, iii) you see motion or hear barking, but in a different direction entirely to the one that your pet ran off in. Any of these cases might suggest that the particular observed cue is related to some other animal than yours and should be discounted in your search. This example is a problem of causal structure inference, where the causality of your observation (“did this observation indeed come from my pet, or some other pet?”) is uncertain and to be computed. These two problems are clearly related in that knowledge or uncertainty about one can create knowledge or uncertainty about the other.

We give an example of these types of multisensory perception problems from three synergistic perspectives.

1. Theoretical modeling of optimal solutions to simple instances of these problems.

2. Building artificial cognitive systems which learn to solve real instances of these problems using machine learning techniques.

3. Investigating natural cognitive systems (humans) to discover how they solve these problems.

Each of these approaches is briefly described in the following sections.

Research

Theoretical Modeling

An elegant and successful approach to modeling perceptual problems is the “ideal observer” approach [5], which we can formalize using Bayesian networks [1].

1. Firstly, we model the process by which signals from the source are observed by the cognitive system - including any distorting noise processes (e.g. sound waves arriving at your ears with slightly different attenuations and delays depending on the source location).

2. Secondly, we assume that the observer, through evolutionary optimization or learning, knows all the parameters of this process.

3. Finally, we ask what an optimal Bayesian observer equipped with this generating model and its parameters would compute about the source given his current observations.
Figure 1: Bayesian network describing the potential correlations among audio microphone elements $x_1, x_2$ and camera image $y$ as function of source location $l$, audibility $W$ and visibility $Z$.

For multisensory observations, in the absence of uncertainty about causality, the ideal observer's computation is known as sensor fusion. Here, a common parametric form describes observations as linear Gaussian functions of the source state$[2]$. This has two (intuitive) consequences. i) The optimal estimate of the state is given by the precision weighted mean of all the observations, such that more reliable observations are given higher weighting. ii) The addition of any further modalities of observation can only increase the precision of the state inference, which is given by the sum of the precisions for each observation modality. So suppose (plausibly) that visual and auditory localization are similarly precise on a foggy evening, while vision is much more precise than audition during the day and vice-versa during the night. Then, on a foggy evening, the ideal observer would rely on an approximately equally weighted combination of vision and audition. During day or night, the ideal observer would rely almost entirely on vision and audition respectively. In this way using audition and vision together is guaranteed to be helpful in localization.

In the presence of causal uncertainty, the ideal observation must also determine the causality of the observations while optimally estimating the state of interest. (In Bayesian network terminology, these two unknowns are conditionally dependent.) This is known as causal structure inference or data association and is formally a Bayesian model selection problem$[7]$. Assuming again that observations are linear Gaussian functions of the state, this model selection will depend on two things: The agreement between the observations as well and the match between the observations and their known statistics. (E.g. to optimally locate the lost pet, the identity of an observed bark also needs to be determined. This will in turn depend on the match between the bearing of the bark and the last known pet location as well as the tone of the bark and the known tone of the pet).

Applying this theoretical modeling approach, we aim to build artificial cognitive systems which are as close as possible to the ideal observer given the available computational resources. (Exact ideal observer inference may be intractable). We can also investigate how closely the perceptual performance of humans comes to that of the ideal observer.

**Artificial Cognitive Systems**

An example machine learning system for Bayesian multisensory perception$[3, 4]$ learns to understand office conversation scenarios audio-visually given a computer equipped with two microphones and a camera. This problem is related to the lost dog example insofar as the use of audition and vision for localization, but somewhat more involved.

The computer is equipped with an ideal observer Bayesian network model (Fig. 1) describing sources (e.g. people) in the world as potentially (but not necessarily) audible and/or visible. It also describes the potential correlation in these observations if a source is simultaneously both speaking and visible.
Figure 2: Operation schematic for audio-visual scene understanding system. Using raw data (a) as input for unsupervised learning and inference in the model (Fig. 1) allows tracking (b), appearance learning (c) and speech segmentation (d).
(via the dependence of the inter-microphone signal delay on the source position). No parameters of this model (e.g. what people look or sound like) are pre-specified. Given the chance to observe raw data of some people moving and speaking (Fig. 2a), the system bootstraps itself without supervision, learning the model parameters (e.g. people’s visual appearance, Fig. 2c) from correlations in the data using expectation maximization\cite{1}. It infers in real time the people’s states, including where they are (Fig. 2b) and de-noised estimates of their appearance and speech. Simultaneously, (and purely by inferring the causal structure of the Bayesian network), the model effectively judges at each timestep whether each person is present, whether they are speaking and who said what (Fig. 2b,c). Note that unlike the lost dog example, the causal structure here is of intrinsic interest, as knowing who said what in a conversation may be crucial to its understanding. The ideal observer model effectively judges causal structure questions such as who said what (Fig. 2d) on the same bases described earlier: Whether the visual and auditory cues come from the same bearing, who was last thought to be at that bearing and the match between the visual and auditory cues with the parameters learnt for each person. In addition, the structure inference allows effective tracking through occlusion/silence in the visual/auditory modalities - a typical point of failure in traditional approaches.

Natural Cognitive Systems

The last decade of human multisensory perception research suggests that (without causal uncertainty) perception is very close to optimal for various tasks using many combinations of senses\cite{2} (e.g. audio-visual localization, visual-haptic size perception, etc). This is typically tested by checking that when humans observe targets with additional sensor modalities, the effects predicted by the ideal observer theory are elicited. Specifically, that the final percept is the precision weighted mean of the individual cues, and that the precision of the final percept is the sum of the precisions of the individual modalities.

More recently, with the theoretical models for causal structure inference described here, human multisensory perception is beginning to be seen as close to ideal even under causal uncertainty. In [6], for example, human subjects are presented with a very similar task to the lost dog or speaker localization tasks discussed earlier. Subjects observe a potentially location discrepant audio-visual stimuli which they must localize and judge for common causation or not. Subjects’ perception of the stimuli exhibits the predictions of an optimal observer faced with multisensory observations of uncertain causality: i) Their judgement of common causation depends on the agreement between the stimuli. ii) When the stimuli are nearly in agreement their percept is almost the precision weighted mean of individual stimuli (because the stimuli almost certainly have common cause). When the stimuli are are in strong disagreement, people’s percept is almost entirely independent of one modality (because the stimuli almost certainly do not have common cause).

Conclusions

In summary, we have seen how the ideal observer approach can be used to understand perceptual problems in cognition involving multiple observations and uncertain causal structure. This links three related strands of perceptual research; theoretical modeling, machine perception systems and neuroscience. Recent research has illustrated the importance of optimal computation of causal structure for artificial cognitive systems (e.g. to understand who said what in a conversation) and has begun to show the optimality of human multisensory perception under uncertain causality.

References


