Probabilistic Modelling and Structural Inference for Multimodal Cue Combination

Sethu Vijayakumar & Timothy Hospedales

School of Informatics, University of Edinburgh, Edinburgh, EH9 3JZ, UK

This chapter addresses cue combination from the perspective of probabilistic modeling and machine learning. We introduce probabilistic graphical models (Jordan 1998) as providing a framework for clearly formalizing perceptual problems – an exemplar of which is cue combination and, in the process, understanding the issues faced by the human central nervous system (CNS) in solving such problems. The solution to inference in probabilistic model formalized for a given perceptual problem provides the "ideal observer" strategy against which human experimental performance can be compared. For example, the frequently cited "rule" of inverse variance weighting for cue combination is the solution to inference in a probabilistic model with a single unknown and many noisy Gaussian distribution observations. Various chapters in this book make this comparison in detail for specific combinations of perceptual cues.

In particular, we will illustrate the use of probabilistic graphical models to understand cue combination problems where there is uncertainty about the model that *should* describe the data. This has been called structure inference (Hospedales et al., 2007) or causal inference (Kording and Tenenbaum 2007). Consider the problem of following a multi-party conversation. This ideally requires solution to the problem of combining visual (lip) information with auditory information to best understand and disambiguate speech. However, the CNS must also solve an additional problem of correctly associating the cues, i.e., appropriately matching speech segments with the person who uttered them before combining the cues. Both components of this problems can be understood in a unified way in the probabilistic modeling framework – in machine learning jargon, we have a model selection as well as an inference problem. For example, in a multi-party conversation involving two other participants, it might be necessary to model each of the potential contributors as two different models with different conditional dependencies, representing the hyptheses that the observed speech was uttered by person A or B. In this case, the optimal observer infers both the source of the speech (the model) and the content of the speech (the latent variable). While this may seem initially like a needlessly complex solution, it is actually a principled and a powerful approach. This single framework will turn out to explain a remarkable variety of experimental data (Kording & Tenenbaum 2007).

As specific examples, we consider the following audio-visual (AV) experiments. In the localization experiments of Wallace (Wallace et al., 2004), subjects must localize an audio stimulus in the presence of another visual stimulus which may or may not be spatially conincident. In this case, the model structure uncertainty is whether the stimuli are actually correlated or not, and hence, whether they should be combined or not? Comparing the probabilistic modeling predictions with the experimental results, it turns

out that the CNS does indeed infer the appropriate model as well as the stimulus location on the fly (Kording & Tenenbaum, 2007). In the counting experiments of Shams (Shams & Beierholm 2005), subjects must report the number (count) of AV stimuli presented, in a paradigm where the number presented in each modality may or may not be correlated. The results reflect increased interaction (perceptual fusion) of the number perceived in each modality when the true numbers were similar across modalities, and less interaction when they were more dissimilar. This is again explained by inferring the variables and structure in a set of probabilistic models. When the stimuli are similar, the fused model is more likely, and the cues' estimates are increasingly integrative. Alternately, when the stimuli are more dissimilar, the fused model is less likely and the cues' estimates are increasingly independent.

Thus far the inference about the appropriate model seems to be a necessary nuisance enroute to inferring the latent quantities; however, it turns out to be of intrinsic interest of its own. Returning to the multiparty conversation example, inference of the latent variables in the model may answer the questions of "what was said?" or "where did that come from?" depending on the particular perceptual problem posed. Knowledge of the correct model, however, encodes the causal structure of the data. Inference of the model, therefore, represents the "ideal observer" solution to higher level relational questions such as "who said what".

To further illustrate these points, we describe an unsupervised machine learning application that we have developed (Hospedales et al. 2007) to model audio-visual perception, learning and scene analysis. This model performs cue combination to make inferences about observed humans, including learning their speech, visual appearance and location using unsupervised techniques in addition to inferring the association between cues and sources; allowing the model to understand "who said what" for a realistic, real time video tracking and transcription application.