Representing Uncertainty for Dynamic Multi-Party Social Human-Robot Interaction

Mary Ellen Foster
School of Mathematical and Computer Sciences
Heriot-Watt University
Edinburgh EH14 4AS, Scotland, UK
M.E.Foster@hw.ac.uk

Abstract
We describe how the initial state representation developed for a socially aware interactive robot is being extended to handle uncertainty. It incorporates the full range of information provided by the input sensors, including multiple possible hypotheses, each with an associated confidence value.

Introduction
A crucial aspect of the design of an interactive system is state management: transforming the noisy, continuous hypotheses produced by the low-level input processing components into a form that can be used as the basis for higher-level action selection by a component such as a planner. Intuitively, states represent a point of intersection between low-level sensor data and the high-level structures used for action selection. Since states are induced from the mapping of sensor observations to property values, the challenge of building an effective state manager rests on defining appropriate mapping functions. This process is not always straightforward: often, it is not the sensor data at a single time point that determines the value of a state property, but rather the patterns found in a sequence of signals. Thus, computing the value of a property might combine information from multiple signals and require temporal cross-modal fusion.

A state representation that considers only the highest-confidence input hypotheses is straightforward to maintain and reason with, but discards a great deal of potentially useful information. On the other hand, a representation that takes into account the full set of input hypotheses—along with their estimated confidence scores—can be more robust and informative, but requires more sophisticated methods of maintenance and more complex forms of reasoning.

We are currently addressing this issue in the context of the JAMES robot bartender (Figure 1), which has the goal of supporting socially appropriate multi-party interaction in a bartending scenario. Based on (uncertain) observations about the users in the scene provided by the vision and speech recognition components, the system maintains a model of the social context, and decides on effective and socially appropriate responses in that context. In this paper, we describe how the state representation initially developed for the JAMES robot is being extended to incorporate the full range of information from the input sensors.

Figure 1: The JAMES robot bartender

Representing Uncertainty
Within the JAMES system, the task of the Social State Recogniser (SSR) is to turn the continuous stream of messages produced by the low-level vision and speech recognition components into a discrete representation of the world, the robot, and all entities in the scene, integrating social, interaction-based, and task-based properties. In addition to storing all of the low-level sensor information, the SSR also infers additional relations that are not directly reported by the sensors. For example, it fuses information from vision and speech to determine which user should be assigned to a recognised speech hypothesis, and uses the vision information to estimate whether each customer is currently seeking the bartender’s attention (Foster, Gaschler, and Giuliani 2013).

For the initial version of the SSR, the state was represented as a single list of properties and their values, and a fixed confidence threshold was used to decide whether to include any given property into the state. Details of this representation are given in (Petrick and Foster 2013), and a sample state excerpt is shown in the highlighted portion of Table 1. The state properties are as follows: seeksAttention indicates whether each customer is seeking the bartender’s attention; lastSpeaker, lastEvent, and lastAct contain the recent interaction history; while drinkOrder estimates each customer’s drink order.

This initial representation has since been extended to incorporate two features that the previous version did not exhibit:

- Every property in the state has an associated confidence value, represented as a number between 0 and 1.
- Every property in the state can have multiple values, with each possible value having its own confidence.
Table 1: State excerpt, showing both the previous representation (highlighted portion) and the new representation

Note that this representation is similar to the Discrete distribution used in RDDL (Sanner 2011), the representation language for the recent probabilistic tracks in the International Planning Competition. Table 1 shows a full state using this expanded representation. In addition to the old-style state information in the highlighted portion, this state also includes confidence scores on all properties—meaning that the low-confidence lastAct(A1) relation can now be included—and also shows multiple values for the drinkOrder(A1) relation.

Incorporating multiple hypotheses and confidence scores into the state requires additional processing in the SSR. For the vision data, we make use of the information from the JAMES computer vision system (Pateraki et al. 2013), which provides a continuous estimate of the location, gaze behaviour, and body language of all people in the scene in real time. Every feature reported by the vision system includes an estimated confidence value; these values are incorporated into the state, and also used to determine the confidence value for derived properties such as seeksAttention.

For speech recognition, we make use of the Microsoft Kinect for Windows API (Microsoft Corporation 2013), which produces an n-best list of recognition hypotheses, each with an estimated confidence score, along with an estimate of the sound source angle and the angle confidence. The recognised hypotheses are parsed to extract the syntactic and semantic information using a grammar implemented in OpenCCG (White 2006), while the sound source angle is used together with the location information from vision to estimate which of the customers in the scene is most likely to have been speaking. If a possible speaker is found, the semantic information from speech is used to update the lastAct relation.

In the case that the customer says something regarding their drink order, we also update the value of the drinkOrder relation, using the generic belief tracking procedure proposed by Wang and Lemon (2013), which maintains beliefs over user goals based on a small number of domain-independent rules, using basic probability operations. This allows us to maintain a dynamically-updated list of the possible drink orders made by each customer in the scene, with an associated confidence value for each order.

Action Selection Under Uncertainty

The JAMES system supports two strategies for high-level interaction management: one that employs general-purpose planning techniques in the form of the knowledge-level PKS planner (Petrick and Foster 2013), and one that makes use of a hierarchical MDP model with action selection strategies trained through reinforcement learning (Keizer et al. 2013). Both of these interaction managers were initially designed to make use of the initial state representation, and both are currently being modified to take advantage of the richer state representation described above: the PKS planning domain is being expanded to incorporate additional sensing actions designed specifically to deal with the uncertainty, while the MDP-based interaction manager is being transformed to one that uses POMDPs (Foster, Keizer, and Lemon 2014).

We will shortly carry out user studies to assess the impact of the new state representation and the associated action selection strategies on user interactions with the system, in each case comparing a version of the bartender that deals with all of the above forms of uncertainty to one that does not. Based on the behaviour of previous versions of the bartender—which did not incorporate any of the uncertainty in the state—we expect to see a positive impact in task performance (i.e., the number of drinks correctly served), since the bartender should clarify lower-confidence or ambiguous state hypotheses instead of simply serving what it believes to be the requested drink. On the other hand, it may be that the subjective user judgements will be negatively affected if the system clarifies too frequently in contexts where the original hypothesis is actually the correct one.

References


