Pre-attentive filtering for robot perception of patterned activity

Paul Fitzpatrick

What: The aim of this work is to give a robot the ability to rapidly perceive simple repeated patterns in its sensory input, without prior knowledge of the senses and events involved in the pattern. Such an ability is most useful if it is massively parallel and *pre-attentive* in nature, analogous to early visual processing [4].

Why: Real-time machine perception benefits greatly from heuristics for quickly filtering out irrelevant stimuli and thus focusing computational effort where it is most likely to pay off. The robots built by the Humanoid Robotics Group all use one form or another of such heuristics for visual perception, such as biases towards skin-colored regions, moving objects, and bright stimuli [2]. More recently, we have investigated the utility of periodicity as a perceptual bias, demonstrating cross-modal priming where visually periodic motion influenced the perception of the sound of tools and toys [1]. But are such biases limited to low-level perceptual features? The goal of this work is to apply similar filtering to the detection of patterned activity, both to facilitate autonomous learning and as a building-block for intuitive human-robot communication (see Figure 1).

How: Perception involves many 'missing information' problems which are straightforward to model but difficult to invert. For example, transforming a 3D scene into a 2D view such as our eye might see is a much more tractable mathematical problem than that of recovering the 3D scene given just the 2D view. The basic difficulty is that many possible world states could have produced the same sensory impression, so there is a fundamental ambiguity to contend with. Of course, not all those world states are equally likely to occur, and this fact is explicitly or implicitly used in all computer vision algorithms to generate plausible interpretations of raw sensory input. For our work, which requires real-time parallel processing of images, there is little time to weigh alternative hypotheses – either algorithms must be very simple, or the results must be pre-computed. This work makes use of the second approach, where many possible interpretations of each possible event sequence are considered, and a favored interpretation and measure of confidence is assigned off-line prior to operation.

For an alphabet of k symbols, there are k^n possible sequences of length n. However, if we are concerned only with the *pattern* of symbol recurrence (that is, if we consider a sequence `abbbac` and `zdddza` to be the same pattern), then the number of possibilities is much, much less. The Bell numbers count these:

sequence length n	**distinct sequences (with n symbols)**	**distinct patterns (Bell number)**
5	3,125	52
6	46,656	203
7	823,543	877
8	16,777,216	4,140
9	387,420,489	21,147
10	10,000,000,000	115,975
11	285,311,670,611	678,570
12	8,916,100,448,256	4,213,597

These numbers are important because they suggest that an exhaustive enumeration of patterns (not sequences) is practical, both for off-line evaluation and on-line storage in RAM, for non-trivial pattern lengths.

For short patterns, human expertise can be captured directly by examining the patterns by hand (this was done for patterns of length 5 in less than an hour). Human judgement is particularly important for short patterns since they can be very ambiguous. For longer patterns, an automated evaluation process is preferable, and can be achieved by exhaustively evaluating a set of models, and comparing the probability of the patterns they generate to find the most plausible interpretations.

Human speech	**Human action**	**Robot speech**	**Robot action**
. . .	. . .	. . .	. . .
yellow	[shows yellow toy]	yelo	[looks at toy]
left	[moves toy left]	left	[tracks toy]
green	[shows green toy]	green	[looks at toy]
right	[moves toy right]	right	[tracks toy]
yellow	[shows yellow toy]	yelo	[looks at toy]
left	[moves toy left]	left	[tracks toy]
. . .	. . .	. . .	. . .
yellow	[shows yellow toy]	yelo	[looks at toy]
next?		ahm left?	[nods left]
yes!	[moves toy left]	left!	
green	[shows green toy]	green	[looks at toy]
next?		ahm right?	[nods right]
yes!	[moves toy right]	right!	
. . .		. . .	

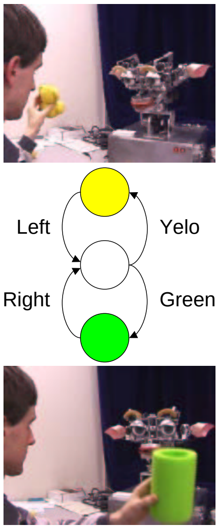

Figure 1: Extracts from a dialogue with a humanoid robot, showing how a simple sorting activity (yellow toys go left, green toys go right) can be verbally communicated to the robot. To the right is shown the simple state machine model of the activity deduced by the robot using a FSM-learning method. The goal of the current work is to facilitate such learning across a much wider range of perceptual inputs than is possible with heavy-weight algorithms.

Progress: Tables have been built automatically for sequences of length up to 10, and by hand for sequences of length 5. Preliminary testing shows results superior to analytic methods previously used (primarily because those methods needed to be weakened to run in real-time, a trade-off not needed for off-line preparation). Some initial experiments have been done with noisy sequences – this requires longer to build tables, but has little impact on run-time operation. Currently the model of activity used is equivalent to regular expressions augmented with the ability to refer to previous sub-expressions. Models are compared based on their description length and specificity.

Future: This work is motivated by recent advances in processor speed and cache size. Much previous work needs to be re-evaluated, to see what algorithms have input spaces that are small enough to allow them to be converted to look-up tables (and 'small enough' can now be quite large!) for fast real-time operation. Of course, this conversion is not always possible, especially if there is significant contextual information that needs to be factored into the interpretive process. But for pre-attentive biases, it seems to make sense.

Research Support: Funds for this project were provided by DARPA as part of the "Natural Tasking of Robots Based on Human Interaction Cues" project under contract number DABT 63-00-C-10102, and by the Nippon Telegraph and Telephone Corporation as part of the NTT/MIT Collaboration Agreement.

References:

[1] A. Arsenio and P. Fitzpatrick. Exploiting cross-modal rhythm for robot perception of objects. In *Proceedings of the Second International Conference on Computational Intelligence, Robotics, and Autonomous Systems*, Singapore, December 2003.

[2] C. Breazeal, A. Edsinger, P. Fitzpatrick, B. Scassellati, and P. Varchavskaia. Social constraints on animate vision. *IEEE Intelligent Systems*, 15:32–37, July/August 2000.

[3] K. Murphy. Learning finite automata. Technical Report 96-04-017, Santa Fe Institute, 1996.

[4] H. C. Nothdurft. The role of features in preattentive vision: Comparison of orientation, motion and color cues. *Vision Research*, 33:1937–1958, 1993.