Multirepresentational Architectures for Human-Level Intelligence:
Papers from the AAAI Fall Symposium (FS-09-05)

Cognitive Architecture and Perceptual Inference

Luis A. Pineda
Computer Science Department, IIMAS
Universidad Nacional Autónoma de México
Ciudad Universitaria, México DF
luis@leibniz.iimas.unam.mx

Abstract

In this position paper we discuss some general properties involved in high-level cognition in situated agents, and sketch an architecture relating sub-symbolic and symbolic information, in which the notion of perceptual inference plays a central role.

An expectation driven architecture for high-level interaction

High-level cognition is embedded in interactive behavior, in which agents relate with the world in "tied" interaction cycles. However, cognition is distinguished from merely reactive behavior in that agents conceptualize or interpret the world in terms of individuals, properties and relations (of a more or less definite character) and this suggests the need for structured representational formats. On the other hand, as cognition is an embedded process, it cannot be detached from perception and action, so the relations between perception and thought (i.e. input) and thought and action (output), need also be considered in a general view of cognition. Perception and action, in turn, relate agents with the external world through sensory and motor capabilities, which have an indeterminate and noise character, and can be better thought of in terms of unstructured or sub-symbolic processes and representations. Accordingly, a general notion of cognition suggests a modular architecture, with structured and non-structured components, where the relation between symbolic and sub-symbolic information plays a central role.

Reactive behavior is to a large extent context independent. Reactive agents are able to sense a number of external stimulus types, and have a set of action types that enable them to respond automatically to specific stimulus. Cognition, on the other hand, is contextually dependent. Agents perform interpretations and actions in contextually rich situations. The context places agents in relation to space and time, and also in relation to other agents with differentiated roles and goals; the context also involves a task-oriented domain, general and domain specific knowledge, and the expectations of agents in specific interaction situations. Consequently, cognitive agents rely heavily on a representation of the interpretation context. Furthermore, cognitive agents represent interpretations of the world rather than the world directly (i.e. a representation is an interpretation). Here, we refer to the process that produces or synthesizes such interpretations as "perceptual inference", and pose that this process plays a central role in cognitive architectures.

Perceptual inference relies on modality specific "images" produced through recognition processes (bottom-up mainly), and the inference itself assigns an interpretation to such images in relation to the context. This process may be mediated by modality specific memories, which are related to (indexed by) the expected intentions and actions in the interpretation context. Interpretations are rendered as descriptions, and these are the objects that are given to cognition proper. In summary, interpretations depend on both the hypotheses provided by modality specific sensory capabilities, on the one hand, and the agent's knowledge about the context and a priori knowledge about the world, on the other. This suggests a "Bayesian" model of interpretation in which the likelihoods depend on sub-symbolic information processes but a priori knowledge has a symbolic or structured representation, and the operator relating these two kinds of knowledge is the interpretation process itself, which we call perceptual inference.

These general observations have inspired a three layer cognitive architecture, as follows:

1. A basic "recognition" layer which translates external information (e.g. light, speech) into sets of features that can be thought of as modality specific "uninterpreted images" (e.g. visual

Copyright © 2009, Association for the Advancement of Artificial Intelligence (www.aaai.org). All rights reserved.

descriptors, text) in a bottom-up fashion (mostly). The object to be recognized may be the external world directly, or an external representation (e.g. diagrams, spoken language or text). It is also considered that external information is noisy and has a continuous and stochastic character, so the construction of faithful concrete "photographic" images of the world, or of external representations, is not feasible most of the time, but not required in general. There is a corresponding output layer that is responsible for performing behaviors directly (e.g. motor behavior, spoken language). Reactive behavior is characterized when uninterpreted images of the world cause output behavior within this layer directly; in this case the interpretation of the stimulus is the corresponding action performed by the agent upon the world.

2. An intermediate layer that performs the perceptual inference proper. This process assigns interpretations to the images produced by the basic recognition layer in terms of the expectations and intentions of the agent in the interpretation situation, and also from modality specific memories. An interpretation may be focused on a single modality specific image or may involve a number of images of different modalities, like visual and linguistic, that may be recognized by the agent at the same time and need to be interpreted in a coordinated fashion. The interpretation of these images depends, in turn, of a qualitative and approximate match between the features produced by the sensory device (i.e. visual descriptors or texts) and modality specific memories, coded or indexed with the same set of feature types, that are relevant to the intentions and expectations of the interpreter agent in the current interpretation situation. The output of this process is a symbolic description expressing an interpretation, and this representational object is the input to the highest cognitive level. The corresponding output layer is responsible of translating abstract specification of actions (e.g. motor or linguistic) into concrete specification of behaviors, in one or several coordinated output modalities.
3. A high-level layer with an explicit representation of the context, where knowledge is organized in structured protocols relating intentions and actions of the agent in task-oriented domains. Specific problem-solving capabilities, like reasoning, planning, theorem-proving, conceptual learning, etc., are all thought of as "abstract actions", that are performed within the context of the interaction protocols. These actions are specified as complex structures and can be partitioned in more simple actions. The interaction protocols are "walked through" hand in hand with the interaction of the agent with the world, so the agent is always situated in the context. Interactive cycles are short, diminishing the need for short-term memory, and preventing that the representations held by the agent diverge from the state of the world. Performing an abstract action at this level may result in the specification of a behavior, which is further determined and executed by the middle and bottom layers, and results on a modality specific action, or a coordinated set of actions, that the agent perform upon the world (or upon an external representation).

A simple version of this architecture has been implemented and tested within the context of a conversational robot, which is able to guide a poster session through a Spanish spoken conversation, coordinated with vision and motor behaviors (Pineda, 2008; Aguilar & Pineda, 2009). The intuitions underlying this architecture have also been considered in diagrammatic reasoning, where the synthesis and proof of geometric theorems with their associated geometric interpretation have been produced through a simple computational machinery once appropriated descriptions expressing the interpretation of diagrams are produced out of external representations (i.e. the relevant diagrams and diagrammatic sequences) through perceptual inferences (Pineda, 2007).

References

Pineda, L. A., Conservation principles and action schemes in the synthesis of geometric concepts, *Artificial Intelligence* (2007), Vol. 171 (4), pp. 197-238.

Pineda, L. A., Specification and Interpretation of Multimodal Dialogue Models for Human-Robot Interaction, in Artificial Intelligence for Humans: Service Robots and Social Modeling, G. Sidorov (Ed.), SMIA, México, pp. 33–50, 2008.

Aguilar, W., Pineda, L. A., Integrating Graph-Based Vision Perception to Spoken Conversation in Human-Robot Interaction, J. Cabestany et al. (Eds.): IWANN 2009, Part I, LNCS 5517, pp. 789–796. Springer-Verlag Berlin Heidelberg, 2009.