Proceedings of the AAAI National Conference on Artificial Intelligence, 2006, pp. 1508-1512.

Multimodal Cognitive Architecture: Making Perception More Central to Intelligent Behavior

B. Chandrasekaran

Laboratory for AI Research
Department of Computer Science & Engineering
The Ohio State University
Columbus, OH 43210
Email: Chandra@cse.ohio-state.edu

Abstract

I propose that the notion of cognitive state be broadened from the current predicate-symbolic, Language-of-Thought framework to a multi-modal one, where perception and kinesthetic modalities *participate* in thinking. In contrast to the roles assigned to perception and motor activities as modules external to central cognition in the currently dominant theories in AI and Cognitive Science, in the proposed approach, central cognition incorporates parts of the perceptual machinery. I motivate and describe the proposal schematically, and describe the implementation of a bi-modal version in which a diagrammatic representation component is added to the cognitive state. The proposal explains our rich multimodal internal experience, and can be a key step in the realization of embodied agents. The proposed multimodal cognitive state can significantly enhance the agent's problem solving.

Cognition, Architecture, Embodiment and Multimodality of Thought

Generality and flexibility are hallmarks of intelligence, and this has led to a search for *cognitive architectures*, exemplified by Soar (Newell, 1990) and ACT-R (Anderson, 1996). Different task-specific cognitive systems may be programmed or modeled by encoding domain- and task-specific knowledge in the architecture. They typically posit a working memory (WM), a long term memory (LTM), mechanisms to retrieve from LTM and place in WM information relevant to the task, mechanisms that help the agent set up and explore a problem space, and mechanisms that enable the agent to learn from experience. Proposals for the specific mechanisms along with the representational formalisms on which they work constitute the architecture designer's theory of cognition. Because of their origin in a certain idealization of human cognition, it is not surprising that Soar and ACT-R are useful both to build AI agents as well to build cognitive models.

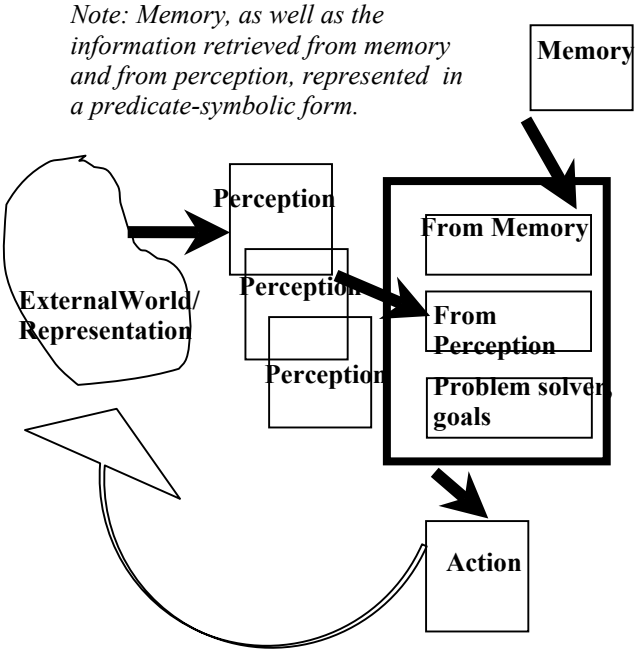

Fig. 1. In the current frameworks, Perception and Action are modules external to Cognition. They do not participate in thinking.

An important aspect of their representational commitment is that the cognitive state, roughly characterized as the content of the WM, is *symbolic*, or to use a more precise term, predicate-symbolic. That is, the knowledge in LTM as well as representations in WM are compositions of symbol strings where the symbols stand for individuals, relations between individuals, or various ways of composing relational predicates, in the domain of interest. For example, in a blocks world, a state representation might be ON(A,B) & Left(B,C). The commitment to symbolic cognitive state representation extends to almost all of AI (knowledge representation) and Cognitive Science (the Language of Thought hypothesis), i.e., is not restricted to the proposals for a general

architecture. Commitment to what I have called predicate-symbolic representation is not limited to logicists in AI. Frame and Script theories, which are proposed as alternatives to logicism, are committed to this form of symbolic representation as the substrate of thought.

The relationship of cognitive architecture as currently conceived to perception and motor systems is given in Figure 1. The rectangle on the right with solid black edges, together with the Memory box, corresponds to an architecture such as Soar or Act-R. Both the information supplied by the perception modules (e.g., On(blockA, blockB)) and the action specification to the Action module (e.g., Move(A, Table)) are in predicate-symbolic form. The perception and action modules help deal with the external world, but they don't do any "thinking." That is performed by cognition using predicate-symbolic representations.

In contrast, consider the phenomenology of our inner selves. We often solve problems imagining spatial situations and performing what feel like internal perceptions, such as in the problem, "Imagine taking a step forward, a step to the right, and a step back. Where are you with respect to the starting point?." Most of us experience "seeing" in an internal image that the final point is one step to the right of the starting point. This phenomenology is independent of the controversy surrounding the "true" nature of mental images. The logic of problem solving is as if a perception is performed on an image. Similarly, a musical composer might solve problems in composition by "hearing" and modifying mental auditory images. In fact, Beethoven is said to have composed a symphony after he became deaf – the problem solving involved in composition must have involved internal auditory images. Deciding if one could make it through a narrow restricted passage, such as a bent tube, requires manipulation of an internal kinesthetic image. In short, in this and similar examples, a perceptual representation, distinct from a predicate-symbolic representation, seems to play a role in thinking, not just providing information about the external world. Robots of the future might take similar advantage of internal images to help their thinking.

Not only is cognitive state multi-modal as described above, memory often needs to support such multi-modality as well. Asked if John was standing closer to Bill than to Stu during an episode the previous night at a party, one might recall an image of their relative locations – something akin to a schematic diagram – and *construct* the answer from the diagram, rather than *retrieve* it from memory. We may not especially have noticed the relative closeness at the time of the episode; thus memory is unlikely to contain a symbolic representation such as Closer-to(John: Bill, Stu). A diagrammatic memory component can support the generation of a large number of predicate-symbolic representations of relations, some unanticipated – e.g., "Was John close enough to Bill to have been able to whisper to him?" – and some not even defined at the time of the episode

Multimodal Cognitive State

What follows is a highly schematic outline of a proposal for a multi-modal cognitive state and associated mechanisms.

To motivate the ideas, let us look at Figure 2. The boxes on the right under the oval together constitute the augmented state and associated systems. For each modality, IPS is the component that supports the internal image. The images in IPS can be created in two ways, by composing elements from memory (as when we imagine "an elephant eating a banana"), and when the agent perceives the external world, i.e., as output of EPS.

The term "image" to refer to the content of IPS may be misleading – they are not the same as the images that are incident at the input of the perceptual modality, e.g., retinal image for vision, or the spatio-temporal pressure waves at the input to the ear. Instead, these are the *outputs* of EPS. This output supports the perceptual experience in the relevant modality, such as the experience of spatially extended shapes in vision, of sounds in the auditory domain, and so on. Operations that result in recognition (categorization) of this experience into an object category ("a peacock") are applied to IPS representations arising from the external world. Relational perception operators (such as "one step to the right of") are applied to the IPS representations – whether they arise from perceiving the external world or from memory-based compositions.

The reader might still be mystified about what makes IPS a category apart from the traditional symbolic representations. The latter involve qualitative abstractions, casting away metric information, and thus cannot support new relational perceptions. We will illustrate the difference with the diagrammatic example, but for now, think of IPS representations as the content of the perceptual experience of a person who is looking at a Henry Moore-like abstract sculpture of shapes, or listening to a cascade of sounds. While this person may have linguistic thoughts associated with his experience, the experience is *not reducible* to his linguistic thoughts. After all, one needs to listen to, not just read about, music to enjoy it. Similarly, the experience is not one of a retinal image of intensities or of pressures on the eardrum – early perception has organized these into perceptual experience of shapes and sounds.

Composability of IPS representations. IPS may also contain images composed from elements from memory, such as when one might imagine a monkey on top of an elephant, never having seen this specific situation in life. Thus, the perceptual representations need be *composable* – in this they are similar to predicate symbol structures. One of the sources of the long-running debate on mental images is the apparent conundrum – reconciling the picture view with the composability requirement. In my opinion, the main reason for the exclusive role for predicate-symbolic in representations as the substrate of cognition in AI and Cognitive Science is the sense that the systematicity and productivity of thought requires compositionality, which

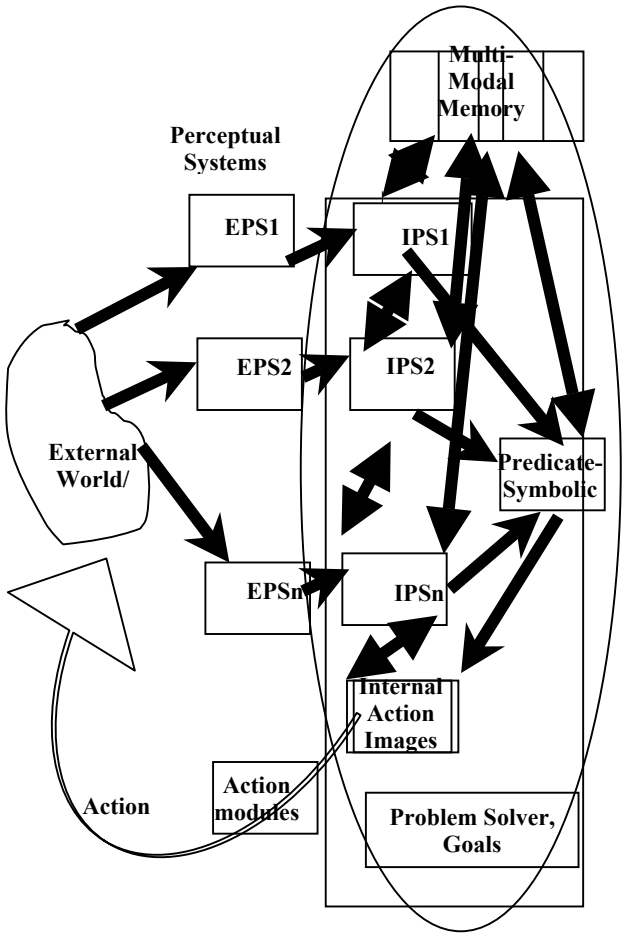

Fig. 2. Schematic of Multi-Modal Cognitive State proposal. The perceptual and kinesthetic modalities are part of the cognitive states as internal images that take part in thinking.

are well-supported by predicate-symbolic representations. However, during the time of the debate, no one had a clue about how images could be composed. Some vision researchers (e.g., Marr (1982)) suggested vocabularies of primitive shapes that could be composed to make complex shapes that retained the full metric properties of 3-d shapes. Chandrasekaran and Narayanan (1990) showed how these proposals could be used to resolve the mental image controversy. As it happens, they are also suggestive of solutions to the representation issues for multimodal cognition. The diagrammatic reasoning architecture discussed below gives an example of a composable perceptual representation. Such composable perceptual primitives can have the requisite metrical information needed for perceptual operations. Additionally, specific representations can be stored in LTM, retrieved and composed, retaining the advantages attributed to predicate-symbol structures

Change of cognitive state. The process of thinking entails changes in cognitive state, in a goal-directed manner. In ACT-R and Soar, cognitive state changes by virtue of rule or operator applications to predicate-symbolic state representations. We may identify this with *inference*, not in the sense of logical inference, but that under the right conditions of matching of information, new information in the form of symbol structures is added. When IPS's are available as cognitive state components, there are additional ways to change cognitive state. One of these is application of a perceptual operator to the contents of an IPS. Thus, if the visual IPS consists of a diagram corresponding to one step forward, one step to the right and one step back, a perception operator can extract the information that the end point is one step to the right of the starting point. This information can be added to the symbolic part, changing cognitive state. Conversely, symbolic contents can create or change an IPS. For example, if we now add the information that the person took one step to the left, the IPS is updated with a new diagrammatic element of a line from the previous end point to the starting point. Performing this modification to IPS would require knowledge in the form of an appropriate diagrammatic element in LTM. The change in the IPS thus changes cognitive state. In general, a change in one of the components can give rise to changes in other components, by *associative evocation*. For example, the word "apple" might evoke the shape of an apple in the visual IPS and a crunching sound in the auditory IPS, and this information could be useful for the next steps in problem solving.

Diagrammatic Representation and Reasoning

To bring this discussion down to earth, let us consider a concrete implementation Unmesh Kurup and I have done of a bimodal architecture where the added modality is diagrammatic.

DRS (Chandrasekaran, 2004) is the name given to the representation framework for representing the diagrammatic component of cognitive state. We consider only diagrams that are configurations of diagrammatic objects each of which is a point, curve, or a region object. Diagrams in DRS are not just arrays of pixels. The diagram in DRS corresponds to the stage in perception where figure-ground separation has been performed, i.e., the image has been organized into diagrammatic objects. DRS consists of a set of internal labels for objects and a complete spatial specification of the objects in a convenient form – the actual underlying formalism is not important, as long as, functionally, the spatial specification is available so that the perceptual routines have the information about the objects that they need to perform their tasks. For example, Inside(point *p*, region *R*) would require the complete spatial specifications of *p* and *R*. Action routines produce diagrammatic objects satisfying given constraints, e.g., the appropriate action routine, when given the constraints, "curve(point *p1*, point *p2*)" and "not-intersect(Region *R*)," will produce a curve from point *p1* to *p2* such that it does not go through *R*. The outputs of the perceptual routines and the constraints for action routines are in predicate-symbolic form, as in the examples given.

A diagram in DRS differs from a physical diagram in that the former is the *intended* diagram: icons and

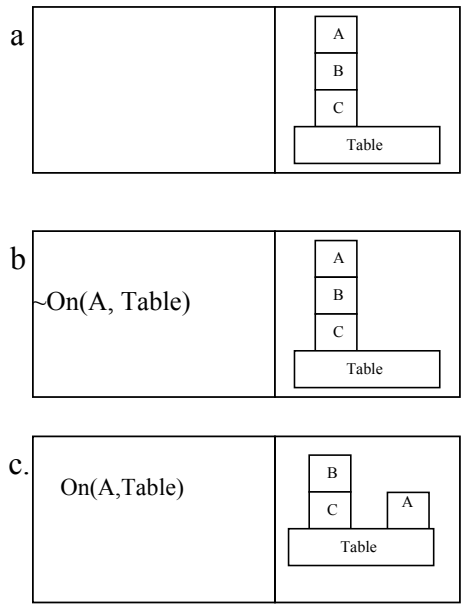

Fig. 3. A series of bimodal states.

alphanumeric characters are abstracted as labels attached to diagrammatic objects, and all objects that necessarily appear in the physical diagram as regions are converted into the intended objects, points, curves, or regions, as the case may be.

Thus, in a Blocks world problem, a state might look as in Figure 3a. Note that in this example only the diagrammatic component contains a representation. Suppose, as part of problem solving, it was necessary to know if A is on the Table. The perception, ON(A,Table), applied to the diagrammatic part will return a negative answer, and the next state would look as in Figure 3b. Suppose now the problem requires that A be on the table, and Move(A, Table) is given as constraint to the action routine, which executes it. The next state might look as in Figure 3c.

An advantage of this representation is that the symbolic component doesn't need to be complete at any point; the information needed can be obtained, if it is available by applying perception to the diagrammatic component. *With respect to the spatial aspects of the problem*, the diagrammatic component is complete in a way that the symbolic component is not, and cannot be. In fact, there is no real reason to carry the complete set of symbolic descriptions from state to state. For situated problem solving, the agent can depend on the external world, and the corresponding internal images in the cognitive state, to significantly reduce the complexity of representation.

DRS is a Symbol Structure *and* a Perceptual Representation. DRS has some of the attractive properties of symbol structures, specifically compositionality. We can imagine Block A on the Table by composing the region object corresponding to Block A with the region object for the Table, and placing the former region above and in touch with the latter region. On the other hand, DRS is not a pure symbol structure with only syntactic relations between them. The region objects, the spatial extents of Block A and the Table, *are* a good part of the semantics of the symbols. The fact that the result of composition is itself a spatially fully specified configuration means that perception operations can be applied on it. In a sense, *we are having our cake and eating it* – the diagram is a composed symbol structure *and* it is an image – resolving the conundrum mentioned earlier.

Learning. In principle, the same learning mechanisms as used in Soar and ACT-R can be used to learn the DRS components as well. The symbol structure corresponding to relevant parts of DRS can be stored in LTM, along with the parameters that can be used to specify the way some primitive shapes may be put together to generate the shape of each object. There are many unsettled issues in the specifics of learning that require further research, but the general forms of the solution are becoming clear.

Other Modalities. DRS provides a feel for the type of representations for other modalities, but first, let us ask what count as modalities. In humans, in addition to sensory modalities such as vision, audition and touch, at least two forms of sensory modality-independent spatial representation exist. The first is egocentric, a sense of space in which we have a more accurate sense of objects near us than of those farther away, something useful for physical navigation without hurting ourselves. The second is an abstract sense of space, such as mental maps that we use to reason about routes. They are sensory modality-independent because vision, audition, touch and kinesthetic modalities can help construct these representations – e.g., humans often use the direction of sound or extend our hands to try to touch nearby objects in the dark, in order to construct a model of the immediate space around them.

How Multimodality Benefits Agents

In situated cognition, access to the external world obviates the need to carry around in one's short-term memory information about the world and reason about changes – to the extent changes are made in the world, the world is its own representation and the consequences can be picked up by perception from the world. This feature of situatedness can be modeled by architectures such as in Fig. 1 – the perception modules can be accessed to get the information. However, when we later need specific information about events we experienced, being able to store the memory in something like a perceptual form, recalling the perceptual form later, applying internal perceptions and answering specific questions can provide economy of storage, since an appropriate perceptual abstraction can stand for a potentially large number of propositions, as discussed earlier. A perceptual representation in memory has the additional advantage that it may be used to answer queries about relations that were defined to the agent after the time of experience – the agent's memory cannot possibly

contain predicate-symbolic representations of such relations.

The real benefits come during reasoning without access to the external world, i.e., reasoning by imagining alternatives in the problem space just as traditional all-symbolic problem solvers do, but where the imagined states have perceptual components. This is what a composer does as she explores the design space of the composition – she needs to experience how a piece of the composition might sound, how a modification of the score might improve it, and so on. A painter has to imagine to some degree the intended painting in his mind's eye. In problem solving involving spatial components, or domain aspects for which spatial analogs exist, the problem solver similarly has to imagine, possibly as a schematic diagram, alternate possibilities, and assessing such states would require applying internal perceptions. Not all this can be done purely symbolically – purely symbolic descriptions of perceptual representations involve qualitative abstractions of quantitative information, and such *qualitative abstractions throw away information that may be needed* for perceptions. Given the locations of three individuals on a surface, no qualitative abstraction of the locations or relations will suffice to answer all the possible questions about the relative locations. If we abstract the original information as, e.g., Left(A,B) & Left (B,C), we won't be able to answer questions such as, "Is A more to the left of B than B is to C?."

The power of human cognition arises at least partly from the seamless integration of language-like thinking based on symbolic abstractions that transcend perceptual modalities, and efficient but modality-restricted perceptual representations and processes. This role of perceptual representations in the process of thinking is also suggestive of the evolutionary development of human-level cognition as built on top of perceptual machinery.

Concluding Remarks

Based on the phenomenology of the content of human thought, I proposed that thinking is not the pure domain of linguistic-symbolic representations and processes, but that perceptual and body representations play more direct roles, in opportunistic collaboration with the symbolic processes, in the production of thought and memory. I proposed that the notion of cognitive state be generalized to a multi-modal representation, in which linguistic symbolic content is one of the "images," along with images in the various perceptual and kinesthetic modalities. I provided some arguments for why and how this might help. I illustrated the proposal by means of our work on a bi-modal architecture involving diagrammatic representations.

Almost all the traditional issues in cognition, problem solving, memory, both episodic and semantic, and learning would be enriched by the multimodal view proposed in this paper. My proposal would be especially useful in the design of robots, where the infrastructure for perception and motor action would already exist, and can be exploited to improve its reasoning capabilities.

Two proposals, one each from psychology and neuroscience, are related to the one in this paper. The first is Barsalou (1999) on perceptual symbol systems. The second is the work by Damasio (1994), in which he locates the basis of thinking on perceptual and body images, which are in turn realized in biological systems as neural activation patterns. Neither of the proposals is, or is intended to be, computational, i.e., unlike my proposal it is hard to directly turn these proposals into AI system implementations. Nevertheless, there are many points of contact and reverberations between my and their proposals.

Acknowledgments

This paper was prepared through participation in the Advanced Decision Architectures Collaborative Technology Alliance sponsored by the U.S. Army Research Laboratory under Cooperative Agreement DAAD19-01-2-0009. I acknowledge Bonny Banerjee and Unmesh Kurup for their contributions in the implementation of the diagrammatic bi-modal version of the architecture, and John Josephson for his ever-available ears for bouncing off ideas, which are usually improved by the said bouncing.

References

Anderson, J.R. and Lebiere, C.J. 1996. *The Atomic Components of Thought*. 1998: Lawrence Erlbaum Associates.

Barsalou, L. W. 1999. Perceptual Symbol Systems. *Behavioral and Brain Sciences* 22: 577-660.

Chandrasekaran, B. and Narayanan, H. R. 1990. Integrating Imagery and Visual Representations. Proc. 12th Ann. Conf of Cognitive Science Society, 670-678. Boston, MA,

Chandrasekaran, B; Kurup, Unmesh; Banerjee, Bonny; Josephson, John R.; and Winkler. Robert. 2004. An Architecture for Problem Solving with Diagrams. In *Diagrammatic Reasoning and Inference*, 151-165. Alan Blackwell, Kim Marriott and Atsushi Shomojima, Editors, Lecture Notes in Artificial Intelligence 2980, Berlin: Springer-Verlag.

Damasio, Antonio R. 1994. *Descartes' Error: Emotion, Reason, and the Human Brain.* Putnam Publishing Group.

Marr, David. 1982. *Vision*, San Francisco: Freeman.

Newell, A. 1990. *Unified Theories of Cognition*. Cambridge, MA: Harvard University Press.