On The Foundations of Perceptual Symbol Systems: Specifying Embodied Representations via Connectionism

Dan W. Joyce[1] (danj@soc.plymouth.ac.uk), Lynn V. Richards[2] (lvrichards@plymouth.ac.uk)
Angelo Cangelosi[1] (acangelosi@plymouth.ac.uk) and Kenny R. Coverntry[2]
(kcoventry@plymouth.ac.uk)
1-Centre for Neural and Adaptive Systems, School of Computing, University of Plymouth. UK. PL4 8AA
2-Centre for Thinking and Language, Dept. of Psychology, University of Plymouth. UK. PL4 8AA

Abstract

Embodied theories of cognition propose that symbol systems are analogue (e.g. Barsalou, 1999; Glenberg, 1997), as opposed to the classicist view that they are *amodal* e.g. Newell and Simon (1976), Fodor (1998). The fundamental problem of symbol grounding (Harnad, 1990) is resolved in embodied theories by admitting only theories of symbolic representation that are grounded in the perceptual system's representation (rather than by reference or mapping of amodal symbols through the sensory systems of the agent). These are often called *analogical* representations (Mandler, 1998). Barsalou's (1999) proposal for perceptual symbol systems (PSS) provides just such a framework for how analogue symbols might come into being, but remains agnostic on the implementation of these PSSs. In this paper, we advance an implementation of PSSs which might fill this explanatory gap. We provide descriptions, an implementation and results from a model and its consequences for Barsalou's theory and embodied representations generally. We constrain our model to the visual modality, but without loss of generality.

1. Introduction

Embodiment theories generally speculate on the overall architecture of cognition (e.g. Clark, 1998; Glenberg, 1997; van Gelder and Port, 1995). Barsalou's (1999) theory moves such theories forward by providing details of how cognitive agents might function with **perceptual symbol systems** (PSSs). In this paper we propose a mechanism for *how* an agent's neural machinery might implement elements of a PSS. Before continuing, we briefly reprise the relevant parts of Barsalou's theory.

A symbol acts as a proximal representation of some distal object for cognitive activities directed towards that object (especially in its absence). The classical view (Fodor, 1998; Newell, 1990, Newell and Simon, 1976) holds that sensory transduction and perception yield a percept which is then transformed into a discrete, amodal symbol – say the token "CAR" for the percept of a car. By *amodal*, it is meant that the symbol or token "CAR" bears no systematic morphological relationship to the percept of a car (that is, the integrated multimodal perceptual experience of a car). The relationship between the symbol ("CAR") and its reference to the natural kind and specific instances (e.g. cars) is the symbol grounding problem (Harnad, 1990). In contrast, Barsalou's (1999) theory of perceptual symbol systems proposes that analogical symbols are those where the representation (i.e. the token) is identified with the perceptual information. There is a systematic, mechanical transformation from the distal object's impression on the sensory surfaces to an internal code (see also, Dretske, 1995). Specifically, Barsalou proposes that:

- perceptual symbols are not recordings, but specific combinations of relevant neural activities induced by current perceptual state -- to include not only visual but all aspects of perceptual experience (Barsalou, 1999; pp 584).
- perceptual symbols extracted for entities (cars) or events (the driving of a car) are collected together in a frame - a structure consisting of many perceptual symbols for the "car" category as it has been experienced in the past
- a simulation is a productive combination of perceptual symbols where a perceptual symbol and its frame relationships are brought into play to produce a potentially infinite set of "concepts" (where a simulation is identified with a concept).

Essentially, PSSs, frames and simulations are abstractions on how experience comes to be a memory and the related structures which appear to support inferences, categories, the support of concepts and so on.

2. Requirements

To proceed, then, we must specify the essential characteristics of a perceptual symbol system's implementation. We will focus on the visual aspect of PSSs. First, an implementation must capture the relevant and salient neural recoding of the sensory information (recall that not everything is captured at any one time in a perceptual symbol). Second, objects (and therefore, events involving objects) rarely exist in temporal isolation and as such, capturing the neural coding must respect the temporal properties of the object (e.g. a symbol is not a static "snapshot" of a

modality). Third, an implementation must provide for categorical representation, e.g. there should be some information in the perceptual symbol that helps it serve as a category exemplar (not necessarily a prototype, but an example of some category). We propose these as at least necessary for the implementation of a perceptual symbol system. We do not claim the following model to be a complete instantiation of Barsalou's theory, but a plausible mechanism for its foundation.

3. Example Application

The network processes labelled video sequences, such as the one shown in Figure 1 (the system is trained on many videos). As a suitable (and plausible) input representation, we propose a "what+where" code (see also Edelman, 2002); that is, the input consists of an array of some 9x12 activations (representing retinotopically organised and isotropic receptive fields) where each activation records some visual stimulus in that area of the visual field. In addition to the "field" representation, we augment a distributed object identity code. These codes were produced by an object representation system (Joyce, Richards, Cangelosi, Coventry, 2002; based on Edelman's (1999) theory) using the same videos. In Figure 2, we show such a coding for the liquid in the video of Figure 1. The distributed object code (bottom left of Figure 2) is appended to each input field.

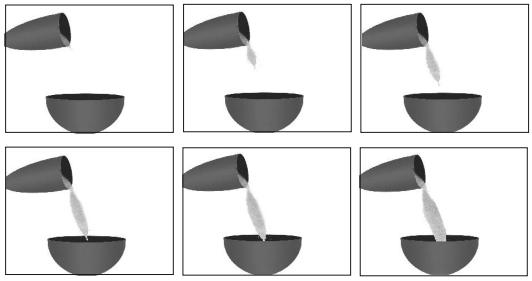

Figure 1 - Example video

4. Connectionist Model

Our model consists of a predictive, time-delay connectionist network similar to Elman's (1990) simple recurrent network (SRN), which we refer to hereafter as the Connectionist Perceptual Symbol System Network (CPSSN). Figure 3 shows the CPSSN network as an Elman SRN, but "folded" about the hidden layer. So, there will be some sequence of neural representations (as described above). The CPSSN is given one set of activations as input which feedforward to the hidden units. In addition, the previous state of the hidden units is fed to the hidden units simultaneously (to provide a temporal context viz. Elman's (1990) SRN model). The hidden units feedforward producing an output which is a *prediction* of the next sequence item. Then, using the *actual* next sequence item, back propagation is used to modify weights (see Figure 3) to account for the error. The actual next sequence item is then used as the new input to predict the subsequent item and so on. Using the coding scheme discussed, we have a total input vector of length 116 (where 8 of these 116 elements code for each object, e.g. liquid, bowl, cup etc.). The output is similarly dimensioned, and there were 10

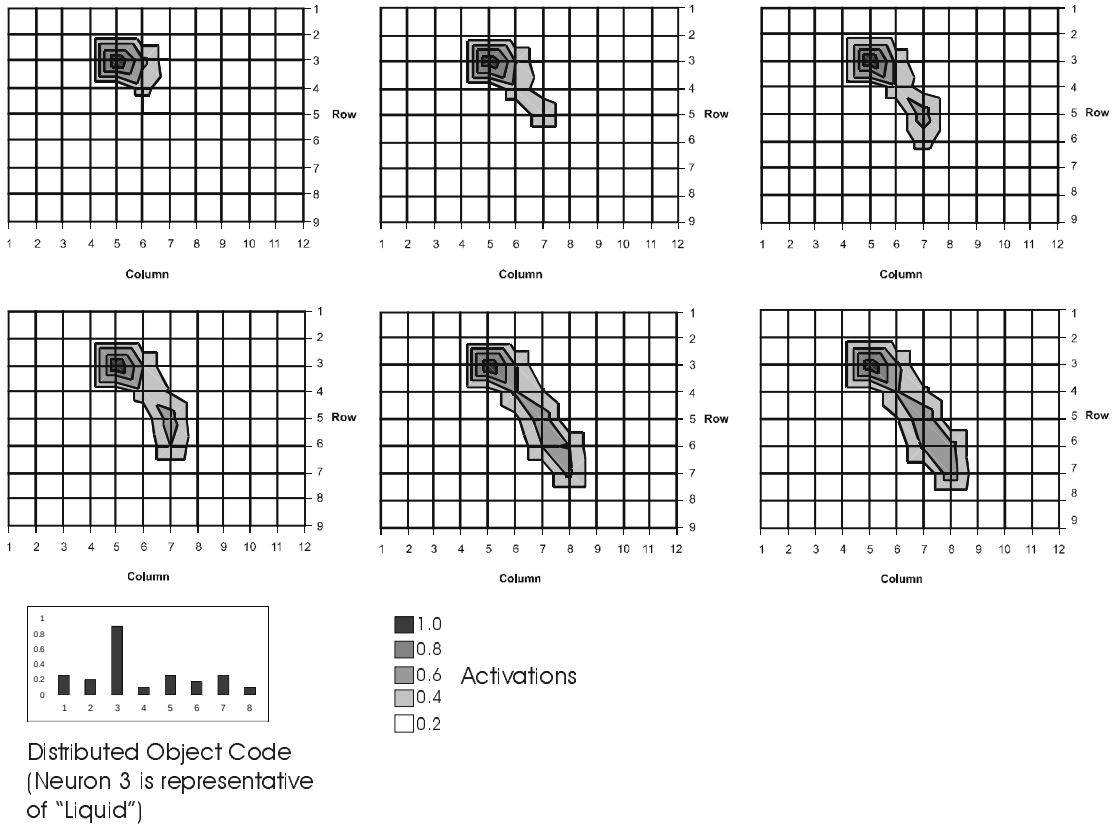

Figure 2 - Example Field Coding for Video (Sequence is presented from top left to bottom right; the reproduced object code is shown in the bottom left, alongside the key showing activation levels. Note that activation of a node is represented at the intersection of the ordinate and abscissa lines)

hidden units (and 10 corresponding time-delayed hidden state nodes).

The network training regime was as follows: a collection of sequences are shown to the network in random order (but of course, the inputs within a sequence are presented one after another). Each sequence contains a field and object code for the "liquid" in the videos. Multiple CPSSN networks would be required to account for the other objects in the scenes. A root-mean-square error measure is used to monitor the network's performance, and the ordering of sequences is changed each time (to prevent destructive interference between the storage of each sequence). Initially, the network is trained with a learning rate of 0.25, and after the RMS error stabilises, this is reduced to 0.05 to allow finer modifications to weights. For 6 sequences, a total of about 150 presentations are required (each sequence is therefore presented 25 times) to reduce RMS averaged over the whole training set from around 35 to around 0.4.

5. Results

It is quite obvious that this network is hetero-associating successive steps in the sequence of fields, but in addition, the network is performing compression and redundancy reduction (in the hidden layer) as well as utilising the state information in the time-delayed state nodes. It is also coding for the changes between sequence items (e.g. the dynamics of how the object moves over time) rather than coding individual sequence items (which would be auto-association). The model embodies the idea that representation is inherently dynamic. The network should, naturally, be able to make a prediction about a sequence given any item in the sequence. Intuitively, the network should be capable of this in the case where a cue is the first item of a sequence, since the time-delayed state is irrelevant (i.e. there can be no temporal context accumulated in the time-delay nodes).

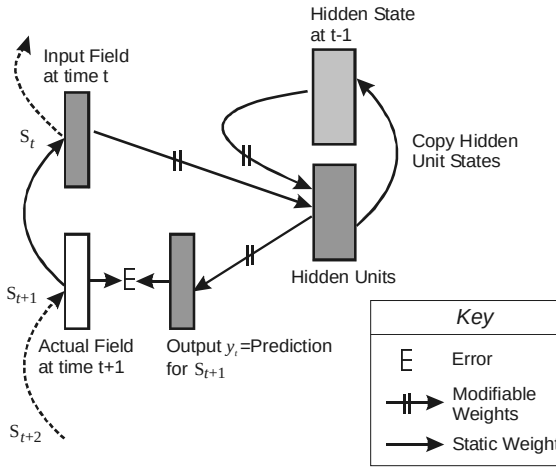

Figure 3 - Schematic of CPSSN Model

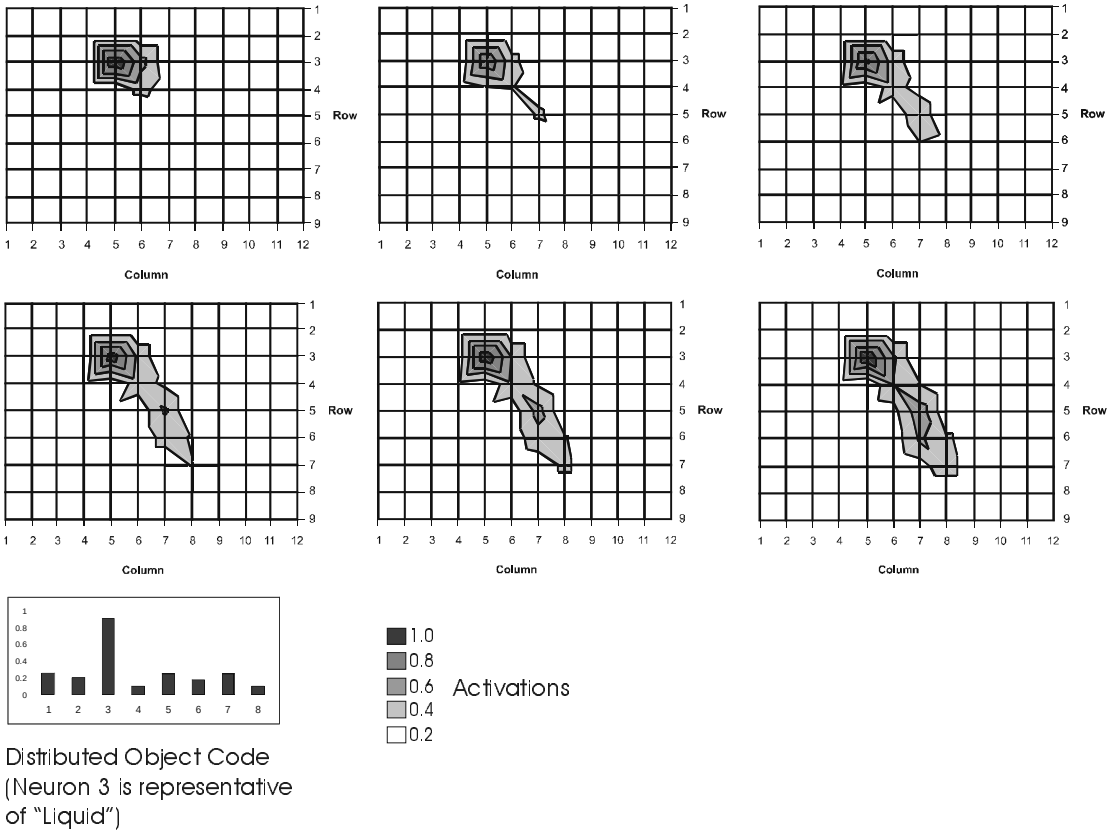

Figure 4 - Sequence Recalled by CPSSN (Sequence reproduced by the network top-left to bottom-right; the reproduced object code is shown in the bottom left, alongside the key showing activation levels. Note that activation of a node is represented at the intersection of the ordinate and abscissa lines)

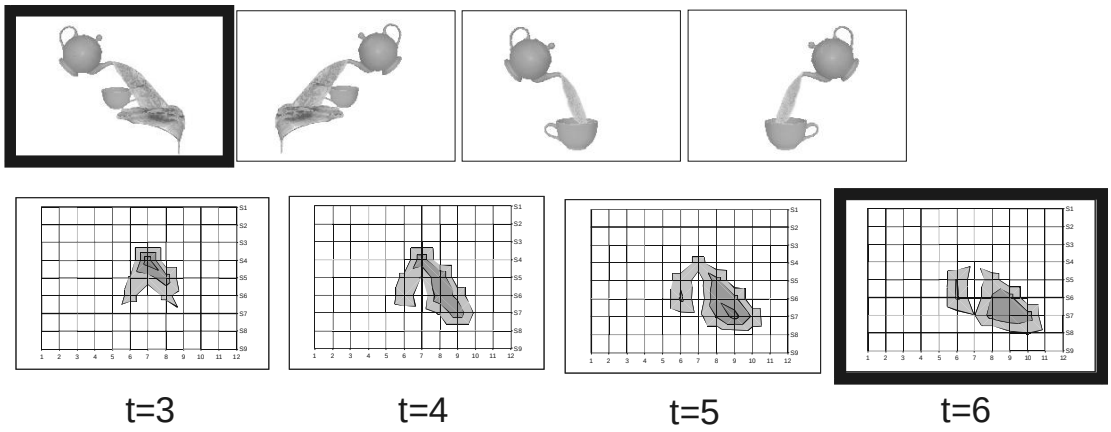

Figure 5 - Sequence Generation with Ambiguity (Same legend as for Figure 4. Top row: the final steps in four videos used in network training for which ambiguity might arise in the early steps of the sequence. Thick black box indicates the video that was used to cue auto-generation shown in the bottom row. Bottom row: last four sequence steps auto-generated by the network. The thick black corresponds to the final state as indicated in the top row. As for Figure 4, only the first step in the sequence was given; the remaining steps are auto-generated)

However, we propose that the network is a mechanism for implementing perceptual symbols, and therefore, a requirement is that it can "replay" the properties of the visual episode that was learned. Given a cue, the network should produce a prediction, which can be fed-back as the next input to produce a sequence of "auto-generated" predictions about a sequence (viz, a perceptual symbol). Figure 4 shows the result of such a test. As can be seen, for the video shown in Figure 1, the movement of the liquid is reproduced with considerable accuracy (compare with the actual training sequence in Figure 2).

However, there are cases in the training data where there is ambiguity that can only be resolved as the sequence evolves and the hidden node states accumulate evidence (that is, as the network produces a number of subsequent, auto-generated steps in the sequence). Figure 5 shows the result of auto-generation when, from a cue early in the sequence, there is ambiguity about which sequence to generate. The top row shows the final step of 4 of the sequences used in training. They have different outcomes (the liquid "splashes" or enters and is contained by the cup) or geometric arrangements, but initially, the flow of liquid will be quite similar. The network reflects this uncertainty (t=3, bottom row) and eventually produces a final activation state which more closely resembles the target.

Examining the evolution of the hidden state reveals an interesting (and useful) property. Again, the auto-generation from first sequence item paradigm was used, and the final hidden-node activation vectors recorded (4 in total, one for each case shown in the top row of Figure 5). 16 pairwise distances were computed, and the resulting matrix of similarities submitted to multi-dimensional scaling, resulting in the 2D projections shown in Figure 6. Notice the linear separability (cf. more complex piecewise discriminants) based on two different categorical features; outcome of the sequence (corresponding to liquid splashes or is contained) and the direction of flow. This suggests that categorical information summarising the event/episode is readily available in the CPSSN (allowing for further recoding for specific cognitive or motoric tasks).

6. Conclusion

It is proposed that the CPSSN model presented captures some of the properties of PSSs. With respect to the criteria from Section 2:

- it captures the **relevant and salient neural recoding** of the sensory information – the input representation captures only relevant neural coding filtered by selective attention (i.e. only the liquid object is processed for the events described; see section 3)
- the neural coding must **respect the temporal properties of the object or event** – in fact, in CPSSN, perceptual symbols are only coded for with respect to the time-evolving dynamics of the visual input. The network can reproduce these (with some uncertainty in ambiguous cases) from a cue (so called auto-generation of the sequence) akin to the properties of Barsalou's (1999) simulations. Our model is also congruent with a dynamical systems theory of embodiment (e.g. van Gelder and Port, 1995; Clark, 1998). The hidden state activations are a context-sensitive representation of the evolving state of the final categorical information as a simulation (e.g. auto-generation from a cue) proceeds.

- an implementation must **provide for categorical representation** – as described above, the system's evolving, cumulative hidden-node state information results in a categorical representation, grounded in the representation (perceptual symbol) and dynamics (the replaying of a sequence from cue) of what was experienced. As Figure 6 shows, a discriminant can be induced that might be relevant for further cognitive processing e.g. inferencing – see Barsalou (1999).

6.1 - Implications for Representation

If Barsalou's PSS theory of embodied representation is implemented in the kinds of computational paradigms we have presented, there are implications for representational kinds; for example, whether space, time or dynamics (e.g. causal effects of forces playing out over time) are the core representational kind upon which others are parasitic e.g. Lakoff and Johnson (1999). Mandler (1998) also suggests that analogical symbols (as we propose the CPSSN model for) hold principle position in the developmental debate: "there need be no propositional representation until language is learned; before that time, infants can use their ability to simplify perceptual input, including objects participating in events, to form analog sketches of what they perceive. These sketches (image-schemas) provide the first meanings that the mind forms" pp. 264. Additionally, Freyd and Finke (1984) defined the term *representational momentum* (RM) to describe the tendency for memory to be distorted in the direction of an implied transformation (another argument for analogical symbol kinds). The CPSSN mechanism *only* codes for changes between sequence items. If such a mechanism underpins perception (and it is intuitive that it might, since the ability to predict and anticipate motion confers a survival advantage – see Hubbard (1998) for discussion), then the kinds of RM effects discovered may be explicable in terms of the CPSSN implementation of PSSs. Such a hypothesis has been recently advanced by Amorim *et al*. (2000) in magnetoencephalography brain imaging studies of RM: "… we hypothesize that RM is a further processing stage that builds on structures devoted to visual motion perception and imagination (mental rotation)" (pp. 578) cf. "replaying" / simulation of events involving perceptual symbols. Futterweit and Beilin (1994) repeated the Freyd study with adults and children, and also found a forward-transformation effect, but only when the photographic images suggested motion. A CPSSN network trained on images showing no motion would likewise not produce a forward-transformation, whereas for motion, it would.

6.2 - Implications for Barsalou's Theory

We propose that CPSSN captures some fundamental aspects of how PSSs might be implemented. While we do not claim that they implement all of Barsalou's proposals, we believe it may provide a foundation. Implications for PSS theory as Barsalou proposes are in how Barsalou views compositionality. In connectionist models, compositionality is of the superpositional, rather than the classical componential kind (van Gelder, 1990). Barsalou appears to favour classical compositionality in his structural descriptions of frames (e.g. he use role-attribute binding of the kind found in prepositional theories). We suggest that this may not be plausible, given the way PSSs are likely to be instantiated.

6.3 – Future Work

We propose that future work should explore more detailed neuroscientific grounding of such a model. While we consider the kinds of recurrency used here as "in paradigm" with current computational theorising in the cognitive neurosciences, more explicit work is required to ground such a model in recent findings. In parallel, we propose exploration of the ability of such a model to accurately replicate experimental findings such as Freyd's RM effects would contribute to the ongoing theoretical debate on the ontology of representational kinds e.g. Mandler (1998).

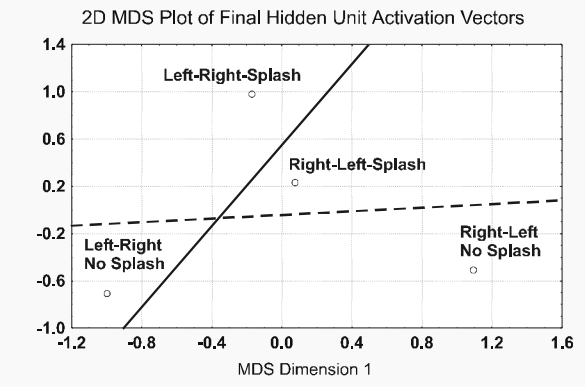

Figure 6 - MDS plot of final hidden unit activation vectors. (Solid line: discriminant for direction of liquid flow; Dashed line: discriminant for outcome of sequence i.e. liquid is contained or splashes)

7. References

Amorim, M.-A., Lang, W., Lindinger, G., Mayer, D., Luder, D. and Berthoz, A. (2002) Modulation of Spatial Orientation Processing by Mental Imagery Instructions: A MEG Study of Representational Momentum. *Journal of Cognitive Neuroscience*, Vol. 12, No. 4, pp. 569-582.

Barsalou, L. (1999) Perceptual symbol systems. *Behavioral and Brain Sciences*, Vol. 22, 577-609.

Clark, A. (1998) *Being There*. MIT Press

Dretske, F. (1995) *Naturalising the Mind*. MIT Press.

Edelman, S. (1998) Representation is representation of similarities. *Behavioural and Brain Sciences*, Vol. 21, No. 4, pp.449-498.

Edelman, S. (1999) *Representation and Recognition in Vision*. MIT Press

Edelman, S. (2002) Constraining the Neural Representation of the Visual World. *Trends in Cognitive Sciences*, Vol. 6, pp. 125-131

Elman, J.L. (1990). Finding structure in time. *Cognitive Science*, Vol. 14, 179-211

Fodor, J.A. (1998) *Concepts*. Oxford University Press.

Fodor, J.A. and McLaughlin, B. (1990) Connectionism and the problem of systematicity: Why Smolensky's solution doesn't work. *Cognition*, Vol 35, No. 2, pp.183-204

Futterweit, L.R. and Beilin, H. (1994) Recognition Memory for Movement in Photographs: A Developmental Study. *Journal of Experimental Child Psychology*, Vol. 57, pp. 163-179.

Freyd, J.J. and Finke, R.A. (1984) Representational Momentum. *Journal of Experimental Psychology: Learning, Memory and Cognition*, Vol. 10, pp.126-132

Glenberg, A.M. (1997) What Memory is For. *Behavioural and Brain Sciences*, Vol. 20, pp. 1-55.

Harnad, S. (1990) The Symbol Grounding Problem. *Physica D*, Vol. 42, pp. 335-346.

Hubbard, T.L. (1998) Representational Momentum and Other Displacements in Memory as Evidence for Non-Conscious Knowledge of Physical Principles. In *Hammeroff, S.R., Kaszniak, A.W. and Scott, A.C. (Eds)* Towards a Science of Consciousness II, MIT Press, pp. 505-512.

Joyce, D.W. and Richards, L.V. and Cangelosi, A. and Coventry, K. (2002) Object Representation By Fragments in the Visual System: A Neurocomputational Model. In *Proceedings of the International Conference of Neural Information Processing (ICONIP)* Singapore, in press.

Lakoff, G and Johnson, M. (1999) *Philosophy in the Flesh*. Basic Books.

Mandler, J. (1998) Representation. In *Damon, W. & Lerner, R. (Eds)* Handbook of Child Psychology, 5th Edition, Volume II (Cognition, Perception and Language), John-Wiley & Sons. Inc, NY.

Newell, A. (1990) *Unified Theories of Cognition*. Harvard Univ. Press.

Newell, A and Simon. H. (1976) Computer science as empirical enquiry. *Communications of the ACM*, Vol. 19, pp. 113—126

Van Gelder, T. (1990) Compositionality: A connectionist variation on a classical theme. *Cognitive Science*, Vol. 14, pp. 355-384

Van Gelder, T. and Port, R.F. (1995) *Mind as Motion: Explorations in the Dynamics of Cognition*. MIT Press.