Thinking With Your Body: Modelling Spatial Biases in Categorization Using a Real Humanoid Robot

Anthony F. Morse (anthony.morse@plymouth.ac.uk)
Tony Belpaeme (tony.belpaeme@plymouth.ac.uk)
Angelo Cangelosi (angelo.cangelosi@plymouth.ac.uk)
Centre for Robotics and Neural Systems, University of Plymouth
Plymouth, Devon, PL4 8AA, UK

Linda B. Smith (smith4@indiana.edu)
Cognitive Development Lab, Indiana University, 1101 East Tenth Street
Bloomington, IN 47405-7007

Abstract

This paper presents a model of sensorimotor learning grounded in the sensory streams of a real humanoid robot (the iCub robot). The robot participates in a replication of two developmental psychology experiments, in which it is shown how spatial cues are sufficient for associating linguistic labels with objects. The robot, using auto-associated self-organizing maps connecting is perceptual input and motor control, produces similar performance and results to human participants. This model confirms the validity of a body centric account of the linking of words to objects as sufficient to account for the spatial biases in learning that these experiments expose.

Keywords: Developmental Robotics; Neural Networks; Sensorimotor; Learning; Spatial Bias; Category Learning.

Introduction

At the heart of all sensorimotor theories of cognition is the claim that perception is to a large degree based upon the use of sensorimotor knowledge in predicting the future sensory consequences of an action, either overtly executed or covertly simulated (Gallese & Lakoff, 2005; Morse, Lowe, & Ziemke, 2008; Noë, 2004, 2009; O'Regan & Noë, 2001). As such our perception of continuous contact with a rich visual world laid out in front of us is somewhat misleading, as sensory input is highly impoverished by comparison to perception; for example visual acuity is focused on an area the size of a thumb nail at arm's length. From a sensorimotor perspective, our perception of things outside the fovea is largely constructed from predictions of what you would see were you to look in this or that direction (Noë, 2004). Clearly such perception is supported by processing of the sparse input from the periphery of our visual field, and mechanisms drawing attention to movement, flashes, and other such changes, yet there remains a large disparity between sensory input and perception.

In taking a sensorimotor perspective, the recognition and categorization of objects in our perceptual field can be achieved through the identification of profiles of interaction unique to each object category. As an example we can perceive a plate as round, not because it projects a round image onto our retina, but rather because we can predict how our sensory contact will change as we move a little this way or a little that way. This rather sparse account supposes that such profiles can be constructed and recognized, leading to the recognition of objects in the world in terms of their Gibsonian affordances (Gibson, 1979). This construction of profiles of interaction is crucial to the ability of sensorimotor theories to account for high-level cognitive and mental phenomena such as perception, but is also the least detailed and most challenging aspect of these theories. Few sensorimotor theories do more than just suppose an ability to do this. Nevertheless such embodiment centric accounts of perception are supported by a large number of psychology experiments and neuroscientific evidence exposing various bodily biases in categorization (Richardson & Kirkham, 2004; Smith, 2005; Smith & Samuelson, 2010). For example, for Gallese and Lakoff (2005) the biological sensorimotor system is not merely foundational to our mental conceptual abilities but constitutes action and perception which are inseparably interwoven in those sensorimotor systems. In addition, the re-activation of visual and motor areas during imagined actions (Jeannerod, 1994; Kosslyn & Press, 1994) "shows that typical human cognitive activities such as visual and motor imagery, far from being of a disembodied, modality-free, and symbolic nature, make use of the activation of sensory-motor brain regions." (Gallese & Lakoff, 2005, p. 465). Similarly while paralysis and neuromuscular blockades do not disrupt conscious thought processes (Topulos, Lansing, & Banzett, 1994), the current activity of the motor cortex is highly influential on both perception and thought. Barsalou et al. (2003) highlight some of the ways in which body posture and action affect perception and cognition; for example, subjects rated cartoons differently when holding a pen between their lips than when holding it between their teeth. The latter triggered the same musculature as smiling, which made the subjects rate the cartoons as funnier, whereas holding the pen between the lips activated the same muscles as frowning and consequently had the opposite effect (Strack, Martin, & Stepper, 1988). Moreover, bodily postures influence the subjects' affective state; e.g., subjects in an upright position experience more pride than subjects in a slumped position. Further compatibility between bodily and cognitive states enhances performance. For instance, several motor

performance compatibility effects have been reported in experiments in which subjects responded faster to 'positive' words (e.g. 'love') than 'negative' words (e.g. 'hate') when asked to pull a lever towards them (Chen & Bargh, 1999).

In the remainder of this paper we describe a developmental robotics (Cangelosi & Riga 2006; Weng et al. 2002) model of a simple sensorimotor system grounded in the sensors and actions of iCub, a child-like humanoid robot. The robot then participates in a psychology experiment highlighting the role of body posture and spatial locations in learning the names of objects. Finally we compare the results of the robot experiments to the data from human child psychology experiments conducted by Smith and Samuelson (Smith & Samuelson, 2010).

The 'Modi' Experiment

In a series of experiments related to Piaget's famous A-not-B error (1963), and derived from experiments by Baldwin (1993), Linda Smith and Larissa Samuelson (Smith & Samuelson, 2010) repeatedly showed children between 18 and 24 months of age two different objects in turn, one consistently presented on the left, and the other consistently presented on the right. Following two presentations of each object, the child's attention is drawn to one of the now empty presentation locations and the linguistic label "modi" is presented. Finally the children are presented with both objects in a new location and asked; "can you find me the modi". Not surprisingly the majority (71%) of the children select the *spatially correlated* object despite the fact that the name was presented in the absence of either object. Varying the experiment to draw the child's attention to the left or right rather than to the specific location that the object, when saying "modi", resulted in a similar performance where 68% of the children selected the spatially linked object. The results of this experiment challenge the popular hypothesis that names are linked to the thing being attended to at the time the name is encountered.

In a follow up experiment, using the same basic procedure, one group of children were presented with only a single object labeled while in sight; a second group were repeatedly presented with a consistent spatial relationship until finally an object is labeled while in sight but in the spatial location where the other object was normally presented. In the control group, where a single object is presented and labeled, 80% correctly picked the labeled object over the previously unencountered object; in the second group (spatial competition) a majority of 60% selected the spatially linked object rather than the object that was actually being attended while labeled. In both experiments changes in posture from sitting to standing disrupted the children's ability to link the absent object to the name through space, while other visual or auditory distracters did not. This is strong evidence challenging the simple hypothesis that names are associated to the thing being attended at the time the name is heard, and strong evidence for the role of the body's momentary disposition in space playing a role in binding objects to names through the expected location of that object.

While several other variations of this experiment have been conducted with children, it is these two versions of the experiment that we have replicated with our robot model.

The Robot Experiments

The 'modi' experiments, though not conclusive, strongly suggest that body posture is central to the linking of linguistic and visual information, especially as large changes in posture such as from sitting to standing disrupt the effect reducing performance in the first experiment to chance levels. In our model this suggestion is taken quite literally, using body posture information as a 'hub' connecting information from other sensory streams in ongoing experience. Connecting information via a 'hub' allows for the spreading of activation via this hub to prime information in one modality from information in another. Furthermore using the body posture as a 'hub' also makes a strong connection to the sensorimotor literature reviewed in the introduction; as actions, here interpreted as changes in body posture, also have the ability to directly prime all the information associated with that new position and hence indicate what the agent would expect to see were it to overtly move to that posture. Such predictive abilities are the foundation of sensorimotor theories.

In this experiment we use the humanoid robotic platform iCub, an open source platform which has been recently developed as a benchmark platform for cognitive robotics experiments (Metta et al., 2008). It has 53 degrees of freedom, allowing experiments on visual, tactile and proprioceptive perception, manipulation and crawling. Initial iCub experiments were carried out in simulation through the open source iCub simulator (Tikhanoff et al. 2008), and then adapted and tested on the physical robot platform.

Grounding information in sensory streams

The information linked via the body-posture hub is the result of processing visual input from the iCub robots cameras, taking the average RGB color of the foveal area and using this as an input to a 2D self-organizing map (SOM) (Kohonen, 1998) described in Equation 1, Equation 2, and Equation 3 below. The SOM provides pattern recognition over the input space preserving input topology while capturing the variance of the data. The body-posture 'hub' similarly used the joint angles of the robot as input to another SOM. Though the iCub robot has 53 degrees of freedom, for simplicity in the experiments detailed herein only 2 degrees from the head (up/down and left/right), and 2 degrees from the eyes (up/down and left/right) were actually used, thus the body map of the iCub robot has 4 inputs, each being the angle of a single joint. Further experiments are underway using a more complex body posture map involving all the degrees of freedom of the iCub robot. Finally, auditory input is abstracted as a collection of explicitly represented 'words', each active only while

hearing that word. In the experiments herein these 'words' are artificially activated, though in related work we are using the open source CMU Sphinx library (http://cmusphinx.org/) to provide voice processing, achieving the same result from genuine auditory input.

Both the color map and the body posture map are initialized using random values in the appropriate sensory ranges with an increased probably of values in the extremes of each range until the SOM's have stabilized. Increasing the probability of extreme values ensures that the resulting stable map fully covers the range of possible input values, without this step mid range values would tend to pull in the extremities of the map resulting in poor coverage.

Equation 1: Initial activation of SOM units

$$A_j = \sqrt{\sum_{i=0}^{i=n} \left(v_i - w_{ij}\right)^2}$$

Where A_j is the resulting activity of each node in the map following a forward pass, v_i is an input, and w_{ij} is the weight between that input and the current node. The winning node is the node with the smallest value for A_i

Equation 2: Final activation of SOM units

$$y_i = e^{\left(\frac{-\beta_i}{2\sqrt{n}}\right)}$$

Where y_i is the final activation of the i^{th} node in the map, β is the distance from node i to the winning unit, and n is the total number of nodes in the map. Note: units not within the neighborhood size are set to zero activation, the neighborhood size and learning rate are monotonically decreased and the map is taken to be stable when the neighborhood size is zero.

Equation 3: Weight changes

$$\Delta w_{ij} = \alpha\left(v_i - w_{ij}\right) y_i$$

Where w_{ij} is the weight between input j and unit i, and α is the learning rate.

The neural model forms the upper tier of a 2 layer subsumption architecture (Brooks, 1986) where the lower tier continuously scans whole images for connected regions of change between temporally contiguous images. The robot is directed to orient with fast eye saccades and slower head turns to position the largest region of change (above a threshold) in the centre of the image. This motion saliency mechanism operates independently from the neural model, generating a motion saliency image driving the motor system. This motion saliency image can be replaced with a color-filtered image to provoke orientation to regions of the image best matching the color primed by the neural model.

Using the model described we then replicated the experimental setup used by Smith and Samuelson (2010), linking the activity of the color map and the auditory words to the body map in real time using positive Hebbian connectivity following Equation 4 below.

Equation 4 Positive Hebbian learning

$$\Delta w_{ij} = \alpha x_i x_j$$

Where w_{ij} is the weight between node j and node i, α is the learning rate (0.01), x_i is the activity of the winning node in one map, and x_j is the winning node in the posture map.

These Hebbian associative connections were then only modified from the current active body posture node. Inhibitory competition between any simultaneously active nodes in the same map provides arbitration between multiple associated nodes resulting in dynamics similar to those expressed in Interactive Activation and Competition (IAC) models which have a long history of use in modeling psychological phenomena (Burton, Bruce, & Hancock, 1999; McClelland & Rumelhart, 1981; Morse, 2003).

As the maps are linked together in real time based on the experiences of the robot (see Figure 1), strong connections between objects typically encountered in particular spatial locations, and hence in similar body postures build up. Similarly, when the word 'modi' is heard, it is also associated with the active body posture node at that time. The relative infrequency of activity in the word nodes compared with continuous activity in the color map is not a problem as competition is between nodes within each map and not between the maps themselves. Finally at the end of the experiment, when the robot is asked to 'find the modi', activity in the 'modi' word node spreads to the associated posture and on to the color map node(s) associated with that posture. The result is to prime particular nodes in the color map, the primed color is then used to filter the whole input

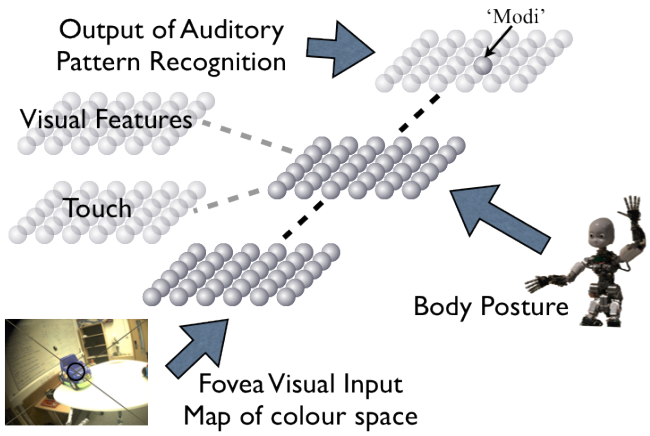

Figure 1: The general architecture of the model. SOMs are used to map the color space, the body posture, and the word space. These maps are then linked using Hebbian learning with the body posture map acting as a central 'hub'. The model can easily be extended to include other features such as visual and touch information in additional SOMs.

image and the robot adjusts its posture to center its vision on the region of the image most closely matching this color. This is achieved using the same mechanism that detects and moves to look at regions of change in the image, replacing the motion saliency image with a color-filtered image. Here the robot moves to look at the brightest region of the color-filtered image, circled in Figure 2 below.

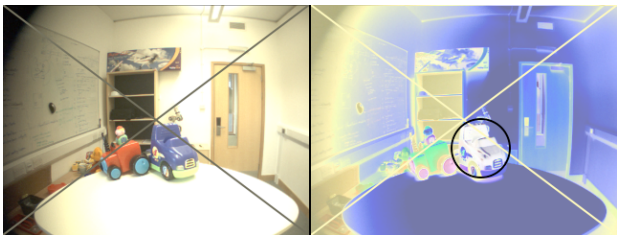

Figure 2 left: Image from the iCub robot's left camera. Right: the same image color-filtered with the primed blue color of the toy truck. The brightest area (circled) indicates the closest match to the primed color.

Given that the number of associations constructed will grow over time in the absence of negative Hebbian learning and in a changing environment, large changes in body posture are used to trigger a removal of these associative connections consistent with the eradication of spatial biases in the psychology experiment following changes from sitting to standing. Additionally, external confirmation that the correct object has been selected leads to more permanent connections being constructed either directly between word and color maps or via a second pattern recognition based 'hub'. As these mechanisms are superfluous to the experiments modeled herein their details have been omitted.

The model as described is then used to replicate each condition of the two psychology experiments described in the previous section as detailed below.

Experiment 1 No Switch Condition

1. Object A is presented to the robot's left – the robot then looks at object A,
2. Object B is presented to the robot's right – the robot then looks at object B,
3. Steps 1 and 2 are repeated,
4. The robot's attention is drawn to its left in the absence of objects A and B and the word 'modi' is spoken,
5. Steps 1 and 2 are repeated again,
6. Object A and object B are presented in a new location and the robot is asked 'where is the modi' – the robot then looks at one of the objects.

This experiment was repeated 18 times resetting the model between each run and starting with a different random seed thereby simulating 18 different individuals. The position of object A and object B (to the left and right) was swapped between each trial and the location that the robots attention was drawn to in step 4 was changed between the first 9 and the remaining trials thereby removing any bias favoring one object or one location over the other. The whole experiment was videoed and stills from steps 1, 2, 4 & 6 are shown in Figure 3. The results recorded which object was centered in the robots visual field following step 6.

Experiment 1 Switch Condition

In the switch condition the location of presentation of objects A and B was swapped for the first presentation only of each object (step 1). Subsequent presentations of each object in steps 2 and 5 remained consistent with the original locations in the no switch condition. Again the experiment was repeated, this time 20 times, with the same variations as used in the no switch condition and the results recoded which object if any is centered in the robots visual field following step 6.

Experiment 2 Labeling while in sight – Control Condition

Experiment 2 provides a variation on experiment 1 in which objects are labeled while in sight. In the control condition a single object is presented either to the left or to the right and labeled 'modi' while being attended, the object is then presented in a new location with a second object and the robot is asked to 'find the modi'.

Experiment 2 Labeling while in sight – Switch Condition

1. Object A is presented to the robots left – the robot then looks at object A
2. Object B is presented to the robots right – the robot then looks at object B
3. Steps 1 and 2 are repeated

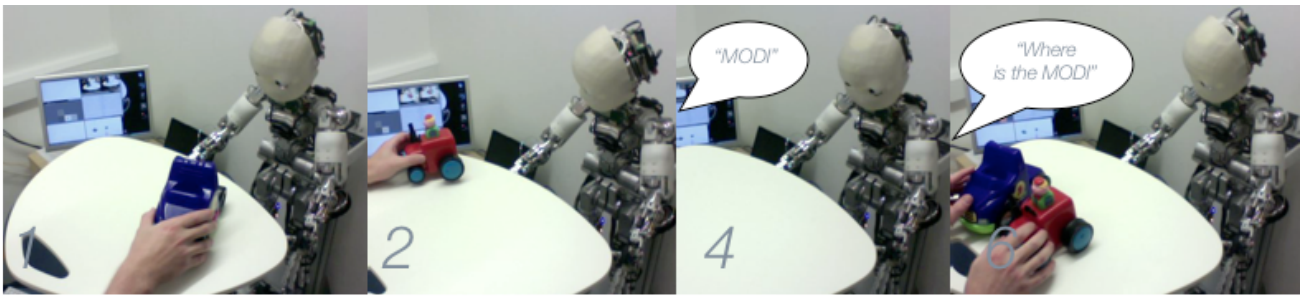

Figure 3: The experiment sequence with the iCub robot.

4. Steps 1 and 2 are repeated again
5. Steps 1 and 2 are repeated yet again
6. Object A is presented to the robots right (i.e. in the wrong location) and the word 'modi' is spoken
7. Steps 1 and 2 are repeated again
8. Object A and object B are presented in a new location and the robot is asked 'where is the modi' – the robot then looks at one of the objects

Experiment 2 was repeated 20 times in each condition with differently seeded networks where the identity of object A and object B was swapped on each consecutive trial and the locations (left and right) were reversed following 10 trials to remove any object or location specific bias.

This model represents preliminary work investigating spatial biases in object categorization. Further work developing and extending this model as a model of sensorimotor learning is currently underway.

Results

In each condition of each experiment, the results recorded which object, if any, was centered in the robots view following the final step of each experiment where the robot was asked to 'find the modi'. In the no-switch condition of experiment 1, 83% (15/18) of the trials resulted in the robot selecting the spatially linked object, while the remaining trials resulted in the robot selecting the non-spatially linked object. This is comparable to the reported result that 71% of children selected the spatially linked object in the human experiment in the same condition (Smith & Samuelson, 2010).

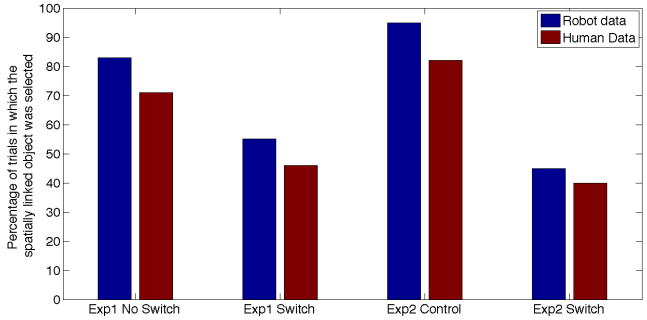

Figure 4: The percentage of spatially linked objects selected in each experimental condition for both robot data and for the human child data.

Reducing the consistency of the object-location correlation in the switch condition resulted in a significant reduction in the spatial priming effect with a close to chance performance of 55% (11/20) of the trials finishing with the spatially correlated object being centered in the view of the robot. The remaining 9 trials resulted in the other object being selected. In experiment 2 objects were labeled while being attended, the control group resulted in 95% (19/20) of the trials selecting the labeled object while in the switch condition only 45% (9/20) of the trials resulted in the labeled object being selected. The remaining trials all selected the other object. These results are compared to the reported human child data in Figure 4.

Discussion and Conclusion

The close match between the results from the robot experiments and the human child results reported by Smith and Samuelson (Smith & Samuelson, 2010) suggests that the hypothesis that body posture is central to early linking of names and object, and can account for the spatial biases exposed by these experiments. What is of relevance here is that the relations between the conditions of each experiment are consistent between the human and robot data, rather than the absolute values achieved. As can be seen from Figure 4 the robot data consistently produced a slightly stronger bias toward the spatially linked objects than the human data.

That the priming effect did not cause the robot to always select the spatially linked object in every variation of the experiments was due to a variety of factors including; noise in the input sensors, varying lighting and reflectance properties as objects are rotated slightly, inaccuracies in the orienting mechanism and so on. In combination these factors produced variations in which a node in the color map was activated as one particular object is being observed, this can lead to weak connections between several similar nodes rather than a single strong connection to one node. In the switch condition of experiment 1, this situation more frequently resulted in object B having a stronger connection to the body posture in which object A was more frequently observed, thus object B was more strongly primed and selected. In these cases increasing the consistency in which an object is seen in the labeled location promotes the strengthening of connections leading to that object being selected, as is seen in the no-switch condition of exp. 1.

It is anticipated that the inclusion of other visual features, though likely to be subject to similar variance, would increase the discrepancy between the data from this model and the human data. This would be due to activation spreading between maps, influencing the priming in much the same way a localist IAC model (Burton et al., 1999; McClelland & Rumelhart, 1981; Morse, 2003). Despite this the relative effects of the various conditions across each experiment should remain relatively consistent. We suggest that the close fit to human data could be misleading, as by comparison in the human case spatial priming would be in competition with far more complex factors influencing the saliency of the objects, factors we have not attempted to model here. Conversely such competition may in fact reduce the models tendency to over perform thereby more closely matching the human data.

As indicated in the introduction our model is consistent with the sensorimotor approach to understanding cognition as the model is able to predict the sensory input it would receive were it to move to different body-postures. This information is accessed simply by a spread of activation from primed body-posture nodes in the 'hub'. The model is also easily scaled up to include additional information presented in additional maps retaining the current IAC-like

architecture. Such models are also suitable for use in hierarchies providing a better fit to the underlying biology.

In conclusion our model accurately reproduces the human data from Smith and Samuelson's (2010) experiments, in an ongoing embodied human robot interaction. In fact, the close fit between our data and the reported human data is in part due to the difficulties and inaccuracies inherent in conducting experiments with complex real robots rather than simulations. In future work we are developing and demonstrating this architecture in a variety of related sensorimotor and psychological tasks involving object manipulations. The goal is close empirical studies of robots and children – in which robot models generate new predictions tested in children. Such joint studies should advance robotics, our understanding of human cognitive development, and the nature of embodied intelligence more generally.

Acknowledgements

This work has been supported by the EU FP7 ITALK project (no. 214668).

References

Baldwin, D.A. (1993) Early referential understanding: infants' ability to recognize referential acts for what the are. *Developmental Psychology, 29,* 832-43.

Barsalou, L. W., Niedenthal, P. M., Barbey, A. K., & Ruppert, J. A. (2003). Social embodiment. *Psychology of Learning and Motivation: Advances in Research and Theory, 43*, 43-92.

Brooks, R. A. (1986). A robust layered control system for a mobile robot. *IEEE Journal of Robotics and Automation, 2*(1), 14-23.

Burton, A. M., Bruce, V., & Hancock, P. J. B. (1999). From pixels to people: A model of familiar face recognition. *Cognitive Science, 23*(1), 1-31.

Cangelosi A., & Riga, T. (2006). An embodied model for sensorimotor grounding and grounding transfer: Experiments with epigenetic robots, Cognitive Science, 30(4), 673-689.

Chen, M., & Bargh, J. A. (1999). Consequences of automatic evaluation: Immediate behavioral predispositions to approach or avoid the stimulus. *Personality and Social Psychology Bulletin, 25*(2), 215.

Gallese, V., & Lakoff, G. (2005). The brain's concepts: The role of the sensory-motor system in reason and language. *Cognitive Neuropsychology, 22*, 455-479.

Gibson, J. J. (1979). *The Ecological Approach to Visual Perception*. Boston: Houghton Mifflin.

Jeannerod, M. (1994). The representing brain: Neural correlates of motor intention and imagery. *Behavioral and Brain Sciences, 17*(2), 187-201.

Kohonen, T. (1998). The self-organizing map. *Neurocomputing, 21*(1-3), 1-6.

Kosslyn, S. M., & Press, M. I. T. (1994). *Image and brain: The resolution of the imagery debate*: MIT Press.

McClelland, J. L., & Rumelhart, D. E. (1981). An interactive activation model of context effects in letter perception: Part 1. An account of basic findings. *Psychological Review, 88*, 375407.

Metta G., Sandini G., Vernon D., Natale L., & Nori F. (2008). The iCub humanoid robot: an open platform for research in embodied cognition. In R. Madhavan & E.R. Messina (Eds.), Proceedings of IEEE Workshop on Performance Metrics for Intelligent Systems Workshop (PerMIS'08). Washington, D.C.

Morse, A. F. (2003). *Autonomous Generation of Burton's IAC Cognitive Models*. Paper presented at the EuroCogSci03, The European Cognitive Science Conference.

Morse, A. F., Lowe, R., & Ziemke, T. (2008). *Towards an Enactive Cognitive Architecture.* Paper presented at the International Conference on Cognitive Systems, Karlsruhe, Germany.

Noë, A. (2004). *Action in Perception*. Cambridge, Mass: MIT Press.

Noë, A. (2009). *Out of our heads*. New York: Hill & Wang.

O'Regan, K., & Noë, A. (2001). A sensorimotor account of visual perception and consciousness. *Behavioral and Brain Sciences, 24*, 939-1011.

Piaget, J. (1963). *The origins of intelligence in children*. New York: Norton.

Richardson, D. C., & Kirkham, N. Z. (2004). Multimodal events and moving locations: Eye movements of adults and 6-month-olds reveal dynamic spatial indexing. *Journal of Experimental Psychology: General, 133*, 46-62.

Smith, L. B. (2005). Cognition as a dynamic system: Principles from embodiment. *Developmental Review, 25*(3-4), 278-298.

Smith, L. B., & Samuelson, L. (2010). Objects in Space and Mind: From Reaching to Words. In K. Mix, L. B. Smith & M. Gasser (Eds.), *Thinking Through Space: Spatial Foundations of Language and Cognition*. Oxford, UK.: Oxford University Press.

Strack, F., Martin, L. L., & Stepper, S. (1988). Inhibiting and facilitating conditions of the human smile: A nonobtrusive test of the facial feedback hypothesis. *Journal of Personality and Social Psych, 54*(5), 768-777.

Tikhanoff V, Cangelosi A., Fitzpatrick P., Metta G., Natale L., Nori F. (2008). An open-source simulator for cognitive robotics research: The prototype of the iCub humanoid robot simulator. In R. Madhavan & E.R. Messina (Rds.), *Proceedings of IEEE Workshop on Performance Metrics for Intelligent Systems Workshop (PerMIS'08)*. Washington, D.C.

Topulos, G. P., Lansing, R. W., & Banzett, R. B. (1994). The Experience of Complete Neuromuscular Blockade in Awake Humans. *Survey of Anesthesiology, 38*(03), 133.

Weng J., McClelland J., Pentland A., Sporns O., Stockman I., Sur M, Thelen E. (2001). Autonomous mental development by robots and animals. Science, 291, 599–600.