Cognition for action: an architectural account for "grounded interaction"

Anthony M. Harrison (anthony.harrison@nrl.navy.mil)
J. Gregory Trafton (greg.trafton@nrl.navy.mil)
Naval Research Laboratory
Washington, DC 20375 USA

Abstract

The effects of priming are not limited to semantics but have also been witnessed in visual-motor tasks (Tucker & Ellis, 2001). By generalizing ACT-R's (Anderson, 2007) existing spreading activation account to include visual representations and broadening the context within which associations are established, we have been able to replicate this small but reliable phenomenon both in simulation and embodied on a humanoid robotic platform. This model illustrates that the effect doesn't require strict embodiment (e.g., Barsalou, 1999) but can instead be accounted for with abstract representations that are "grounded by interaction" (Mahon & Caramazza, 2008).

Introduction

One of the current drumbeats in cognitive science is that cognition is for action. The strongest evidence for cognition for action comes from experiments that show that there is a much tighter coupling of perception and action than previously thought. For example, Glenberg and Kaschak (2002) found that when a sentence implied action in one direction (e.g., "Close the drawer"), participants had difficulty making a sensibility judgment that required a response in the opposite direction. Similarly, when participants indicated whether an object like a teapot was upright or upside down, reaction times were fastest when the response hand was the same as the hand that would be used to grasp the object (e.g., the right hand response was fastest if the teapot's handle was on the right) (Tucker & Ellis, 1998). Many of these researchers argue that this data shows that our thinking is fundamentally embodied, not abstract.

The main idea behind the embodied cognition movement is that cognitive representations and operations are firmly grounded in their physical context and that cognition relies heavily on modality-specific systems and actual bodily states (Tucker & Ellis, 1998; Barsalou, 1999; Wilson, 2002; Niedenthal, Barsalou, Winkielman, Krauth-Gruber, & Ric, 2005). The typical counter to embodied cognition theories are old-style abstract/symbolic theories (Newell & Simon, 1972; Pylyshyn, 1984), which argue that actual experience occurs in modality-specific representations, but those modality-specific states are abstracted and preserved as abstract, amodal symbols. Given the strength of abstract/symbolic theories, some have suggested that the only way that these theories can explain embodied effects is by adding increasingly complex post hoc assumptions about representations and processing (Barsalou, 1999; Niedenthal et al., 2005).

Mahon & Caramazza (2008) argue that the strict embodiment argument against abstract/symbolic theories neglects to consider the possibility that activation, spread through abstract symbols to modal representations, can account for these very same phenomena. While recognizing that abstract/symbolic theories could accommodate such tight perceptual/action coupling, they acknowledge that most theories do not adequately specify the computations and representational content that would permit such coupling through the spreading of activation. Such an abstract/symbolic system would be "grounded by interaction" (Mahon & Caramazza, 2008), if the abstract symbols come to be tightly coupled with their related percepts and required actions through experience acting in the environment. In this manner, the system would be able to exploit both the flexibility of the abstract representations and the richer context afforded by grounded representations.

We present an ACT-R (Anderson, 2007) model that fits within Mahon & Caramazza's "grounded interaction" framework (2008) that provides a process explanation of a classic embodied phenomenon – the visual-motor compatibility effect observed by Tucker & Ellis (2001).

Tucker & Ellis (2001)

Tucker & Ellis (2001) report a series of experiments that show a small but significant effect of visual presentation on grasp responses. In experiment 1, participants viewed a series of objects of different categories (e.g., natural or man-made) that were either large or small. The object size maps directly to the normal grasp used to manipulate the object: a power-grip (i.e., full hand) for large objects and a precision-grip (i.e., thumb and forefinger) for small ones. Objects were placed either near the response hand (15cm) or far away (2000cm). Subjects responded with either a power- or precision-grip response based on the category (i.e., natural/man-made) of the object seen. The task response-mapping (e.g., natural/precision) was varied between subjects.

While there were some simple main effects, the critical result from the first experiment was the interaction between the size of the object and response-mapping. Despite the fact that the size of the object was irrelevant to the task, its compatibility with the response-mapping resulted in reduced reaction times and error rates (figures 1 & 2). Specifically, when viewing large objects, power responses were faster and more accurate than precision responses. Similarly, viewing small objects resulted in faster and more accurate precision responses than power responses.

In experiments 2-4b, Tucker & Ellis used a go/no-go paradigm, with the response-mapping cued by a tone and go/no-go cued by the object category. Experiment 2 presented the response-mapping cue tone 500ms *before* object presentation. The lack of a compatibility effect in the results showed that prior knowledge of the required response was sufficient to override the phenomenon.

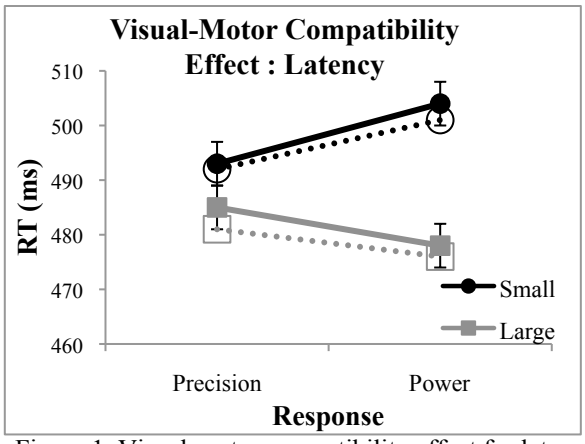

Figure 1. Visual-motor compatibility effect for latency (Tucker & Ellis, 2001, experiment 1). Dotted lines are model fit (R^2=0.99, RMSE=2.95ms).

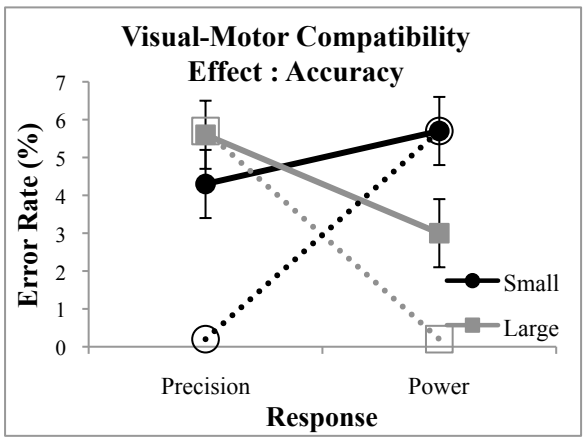

Figure 2. Visual-motor compatibility effect for accuracy (Tucker & Ellis, 2001, experiment 1). Dotted lines are model fit (R^2=0.88, RMSE=2.48%).

Experiment 3 reversed the time delay of the prior experiment and presented the response-mapping cue tone 300ms *after* object presentation. In this circumstance the compatibility effect was present. These results and those from experiment 2 show that the effect is dependent upon the motor system not already being prepared for a particular response.

In experiments 4a and 4b, the visibility of the object was manipulated. In 4a the object disappeared at the same time as the response-mapping cue tone was presented. In 4b the object disappeared 300ms *before* the response-mapping cue tone. The compatibility effect was present in 4a and not 4b, showing that the object's visibility at response selection is critical.

To summarize, Tucker & Ellis have shown that when the object's normal grasp response is compatible with the experiment's response-mapping, there is a small but significant benefit (experiment 1). However, this is conditional on the motor system not already being prepared for a particular response (experiments 2 & 3) and that the object is visible at response selection (experiments 4a & 4b). They discount the theory that this is an example of the percept *directly* priming a particular motor response, arguing that the object would have to be within reach and that such a mechanism would not work for images of objects as well (Tucker & Ellis, 1998). Instead they propose that this is evidence of "a more general representational mechanism that describe object properties in motor terms" (Tucker & Ellis, 2001).

Architectural Account

Within the ACT-R cognitive architecture (Anderson, 2007), the time it takes to retrieve a specific memory (i.e., chunk) is inversely related to that chunk's activation. The chunk's activation is composed of three primary components: base-level activation, spreading activation, and some stochastic noise. Base-level activation is a learned quantity subject to decay that incorporates the effects of frequency and recency of the memory's use. Spreading activation is context dependent, allowing chunks that are the focus of attention to activate related memories. In this way, the chunks within a given buffer (i.e., the focus of attention for a given module in ACT-R) can make related concepts more readily retrievable. Spreading activation is the mechanism used to account for semantic priming effects (Anderson & Reder, 1999). This same mechanism is used here to model the visual-motor priming reported by Tucker & Ellis (2001).

ACT-R defines the current context as the contents of the chunks currently in the model's buffers. If chunk *i* is in a buffer *k*, then all of the chunks that *i* references are in the context. The source activation of buffer *k* is shared equally among those context chunks, and they in turn spread that activation to all the chunks that contain references to them. ACT-R only establishes associative links from the referenced chunk to the referring chunk. The more chunks that reference a specific chunk *j*, the weaker its associative strengths are to the referring chunks. Chunk *j* becomes a less effective retrieval cue because the weaker associative links spread less of the source activation.

This mechanism of spreading activation through associative links from the currently defined context allows ACT-R to model semantic priming (Anderson & Reder, 1999). However, in order to address the visual-motor priming shown in Tucker & Ellis (2001), ACT-R's existing mechanisms must to be modified slightly. These modifications are not complex post-hoc assumptions, rather they are consistent with the existing framework.

Visual Representation and Activation Normally, ACT-R models use only the intentionality system (i.e., goal buffer) as a source of activation, even though all buffers have the capability. Obviously, in order to support visual priming,

the visual buffer must also be used as a source of activation. However, the utility of visual activation is limited due to the traditional structure of the visual representations. The visual representation does not represent a semantic concept; rather, it is a raw percept made up predominantly of non-chunk, primitive features (e.g., numbers, strings). They are therefore highly insular, having little connection to other chunks, which dramatically limits the spread of activation. Typically the visual object's *value* slot (usually a string literal) is used to uniquely associate it with the semantic representation of that percept (i.e., its symbol). To allow a visual percept to activate a semantic symbol, as well as other chunks related to that concept, the *value* slot was modified to reference the semantic symbol chunk directly. To access the semantics, a retrieval must still be made, but now the percept itself can prime that retrieval.

Co-occurring Contextualization Canonical ACT-R only establishes associative links between chunks through symbolic references (i.e., chunk *j* can activate chunk *i* since *i* directly references *j* as a slot value). We propose that symbolic links can also occur through co-occurrence. The context within which processing occurs is not limited to the symbolic structure of the chunks currently in buffer, but actually includes the patterns within the processing units (i.e., productions) that execute cognition. If a production matches against both contents of the goal and visual buffers in order to fire, then the contents of those buffers do not define the context independently, but jointly, and should be linked associatively. Because productions can contain perceptual and motor patterns, perception and action can become linked through co-occurrence.

The application of this mechanism is relatively straightforward. Specifically, the semantic symbol of a percept and the motor command used to grasp the object are associated with each other even though neither has a direct symbolic relationship to the other. These associations are learned from the environment as a consequence of attending to an object, considering its meaning, and manipulating it.

In the language of Mahon & Carmazza (2008), the semantic symbol that is linked to a percept provides the abstract representation that mediates perceptual processing and motor activation. The motor and visual representations are grounded to this abstraction through a history of interaction, allowing the establishment and strengthening of associative links through co-occurrence. Activating the abstract symbol propagates activation both to experienced percepts and motor commands.

Model Details & Fit

The model presented here focuses on a simplification of Tucker and Ellis' (2001) first experiment; how it accounts for the subsequent experiments will be saved for the discussion. Because their presentation distance manipulation had no influence on the visual-motor compatibility effect, it was eliminated from the simulation. Otherwise the simulation is identical to the actual experiment including the timing of object presentations.

Execution The model completed 160 trials (as did participants) where it was presented small and large objects that were either natural or man-made (e.g., strawberry, key, potato, frying pan). Retrieving the visual-symbol associated with the percept, the model was able to classify the object. With this information the model retrieved the appropriate response-mapping for the classification (e.g., natural/precision or man-made/precision). The final retrieval was of the appropriate grip command itself. Once the motor command was retrieved it was passed to the motor system to be executed as the trial response.

Assumptions This model relies upon three key assumptions. First, that activation is spread through not only the goal buffer but also the visual buffer. Second, that the process of encoding a visual percept includes linking that percept to its semantic representation (i.e., its visual-symbol). Finally, associative links are not limited to containment relationships. Over the history of interacting with the environment, both the visual-symbol for a percept and the motor command used to manipulate the object come to be associated with each other via co-occurrence.

Since priming within ACT-R is function of spreading-activation, base-level activations are not of theoretical interest. However, these values do come into play with respect to the rapid retrieval times in the data (figure 1) and are discussed in detail later.

Spreading Activation The model proposes that the visual-motor compatibility effect reported in Tucker & Ellis (2001) is due to activation spreading both from the intentionality (i.e., goal) and visual systems. Once the object is visually encoded, activation is spread to the learned motor response through the co-occurrence associative link between it and the visual-symbol. When the model has retrieved the appropriate response-mapping for the object's category, activation is spread to the task appropriate response. For incompatible responses, activation is spread to two different motor commands. However, when the responses are compatible, both activation sources converge on a single motor command (figure 3). Because of the higher total activation of the compatible motor response, it can be retrieved faster. The lower activation of the incompatible response also makes misretrieval more likely since noise might exceed the differences due to spreading activation.

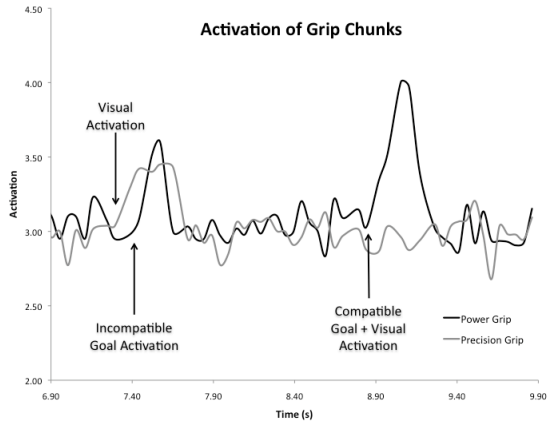

Figure 3. Total activation of incompatible and compatible response motor commands (with noise).

Results and Fits 1000 iterations of the model were run on the simulated version of the experiment. Reaction time fits were quantitatively very strong (R^2=0.99, RMSE=2.95ms, figure 1). Accuracy fits were less strong, but captured the qualitative effect (R^2=0.88, RMSE=2.48%, figure 2). The weaker accuracy fit was due largely to the exclusion of base-level learning from the model. With only spreading-activation, compatible response trials are effectively immune to noise making false retrievals impossible (figure 2 & 3).

Parameters The fits reported above required the manipulation of a few parameters, some of which were dictated by the architecture and the structure of the model.

The maximum associative strength was set to 3.1 (default of 1). This parameter is completely constrained by the structure and connectivity of the model. Conceptually, any chunk that has many references should be a weak retrieval cue; that is its associative strength should be near zero. A maximum associative strength of less than 3.1 would result in negative (i.e., inhibitory) associative strengths for the most heavily referenced chunks. If one of these chunks were to be used as a retrieval cue, it would actually become harder to retrieve its related concept.

The source activation from the goal buffer was kept at the default value of 1. The activation from the visual buffer was set to 0.3, instead of the default of 0. This allows the contents of the visual buffer (namely the semantic symbol) to weakly prime the normal motor response for the object.

While base-level learning was not used in this model, base-level activations were still critical to achieve the rapid retrieval times implied by the average response time of 490ms. Three separate factors influenced the selection of the base-level activations for the visual-symbol, grasp-command, and response-mapping chunks. First, the model, as implemented, requires five productions with three retrievals before a response can be started. At 50ms per production, an additional 85ms for the visual object encoding, and a minimum motor execution time of 50ms, there is only 105ms left for the three retrievals. Second, we assume that over a lifetime of observing and interacting with these objects, the base-level activations for the visual-symbols and grasp-commands are both stable (i.e. relatively immune to decay) and strong. Since we generally see objects more often than we grasp them, visual-symbol activations were set greater than the grasp-commands. Finally, while the response-mapping chunks would benefit from recency, their frequency of use would still be small, so base-level activations were set lower than those of the grasp-command chunks. Base-level activations of 5, 3, and 2.25 were used for the visual-symbol, grasp-command, and response-mapping chunks respectively. While these values are necessary for the low RMSE latency fit, the qualitative (R^2) fit is less sensitive to the base-level values.

To account for the errors in performance, the model relied upon misretrievals. This was accomplished by setting the activation noise parameter to 0.06. The qualitative error results are largely unchanged for most published noise values since the noise only affects the incompatible responses (unless noise exceeds the activation spread to the chunks by the visual buffer). The model's fit of the accuracy data is weaker due largely to the simplification of removing base-level learning. Since compatible responses receive all of the spreading-activation, they are effectively immune to noise (figure 3), which results in 100% accuracy for those trials. To achieve the average 3.5% error rate for compatible trials seen in the data, base-level learning would have to be enabled. This could produce situations where successive retrievals of one particular response might boost its base-level activation such that it could falsely intrude on a subsequent trial. Attempting to fit the error data in this manner would have required seven additional parameters (base-level learning rate, and average age and access counts for the visual-symbol, grasp-command, and response-mapping chunks) instead of the three fixed base-levels used.

Robotic Embodiment

One of the challenges in modeling embodied cognition is the lack of a physical body. This lack is especially relevant because one of the embodied cognition claims is that the body is central to both perception and action; it is disingenuous to claim that we can account for embodied cognition phenomena without a body.

One aspect of running cognitive models on embodied platforms is that actual perception and action must occur. Critically, both perception and action must use cognitively plausible representations and cause the physical body to move.

We have modified ACT-R by allowing it to perceive and act on the physical world by attaching robotic sensors and effectors to it; we call our system ACT-R/E (*Embodied*) (Trafton, Harrison, Fransen, & Bugaska, 2009). Changes to the visual and motor modules are described below.

The Visual Module is used to provide a model with information about what can be seen in the current environment. ACT-R normally sees information presented

on a computer monitor. We modified the original visual module to accept input from a video camera. The visual module allows access to object identification through fiducial (Kato, Billinghurst, Poupyrev, Imamoto, & Tachibana, 2000) and face (Fransen, Hebst, Harrison, & Trafton, 2009) trackers.

Traditional ACT-R has a virtual motor system that allows virtual hand movements (e.g., typing, mouse movements). ACT-R/E's motor module allows control over all of the robot's effectors. When a motor chunk enters into the motor module, a specified motor controller executes the actual physical response.

Our current robot platform is the Mobile-Dexterous-Social (MDS) Robot (Breazeal, 2009). The MDS robot neck has 18 degrees-of-freedom (DoF) for the head, neck and eyes allowing the robot to look at various locations in 3D space and 11 DoF on its four-fingered hand, allowing it to make various gestures and grips. Perceptual inputs include two color video cameras and a SR3000 camera to provide depth information. For the current project, the MDS head can identify various objects through the fiducial tracker and can move its hands in a power or precision grasp.

The 10ms visual-motor compatibility effect is completely obscured by the robot's motor system's slower execution times. In order to illustrate the effect, we dramatically increased the retrieval time scalar. In the video (see acknowledgments for the URL) the visual-motor priming accounts for around a 500 ms performance improvement.

Discussion

ACT-R has a long history of accounting for semantic priming effects (Anderson, 1974), but its perceptual/motor integration has been less explored. To address this theoretical gap, we have modified the visual representation linking the percept to its derived abstract symbol. This allows source activation to usefully spread from the visual system, instead of just from the intentionality system (i.e., goal buffer). We also present a broadened definition of predictive context for the establishment of associative links. Traditional ACT-R only establishes associative links from the contained chunk to the container. In this way, when another chunk has a reference to the contained chunk, it is potentially predictive of the need for the containing chunk. We augment spreading activation to deal with co-occurrence so that we can establish a richer context. In this manner, the visual-symbols and motor commands come to be associated as productions fire that simultaneously match both of the representations in their respective buffers. While only the consequences of this mechanism are exploited in the model presented, the actual process is under active investigation.

A particular limitation of this account is that it does depend upon both visual and motor experience with a given object. Lacking such experience, the modal representations will not be associatively linked to the abstract symbolic representation. As such this model cannot account for related effects when novel objects are used; such as those seen when subjects concurrently perform a compatible manual rotation during the classic Shepard & Metzler (1971) mental rotation task (Wexler, Kosslyn, & Berthoz, 1998).

Experiments 2-4 While the model presented only addresses Tucker & Ellis's (2001) first experiment, its extension to the other experiments is fairly straightforward. All of the subsequent experiments used a go/no-go paradigm where the response to be given was cued by a tone and the go/no-go was determined by the object's category. Recall that in experiment 2, subjects heard the response cue 500ms before the object was presented. This 500ms window of time would allow the model to retrieve the appropriate motor response before it had to determine whether or not to execute it. The lack of a visual-motor compatibility effect observed would be due to the fact that the response had already been selected, leaving visual priming no opportunity to influence performance.

In contrast, in experiment 3 the response cue tone was presented 300ms *after* object presentation. As in experiment 1, the visual presence of the object would allow activation to spread to the learned motor response, facilitating retrieval when it was compatible with the response cued by the tone.

Experiment 4a removed the object at the same time as the cue tone was presented. If the model were able to retrieve the motor command at the moment of the cue-onset and visual-offset, the compatibility effect would be observed. However, ACT-R's encoding time for auditory information would actually result in the retrieval starting at least 50ms after presentation. Since ACT-R's spreading activation mechanism is instantaneous, that activation would drop to 0 immediately after the percept disappeared, eliminating visual priming entirely. The results from experiment 4b are more easily accounted for. Since the object was removed 300ms before the cue tone, the activation of the learned motor response would have been eliminated before the retrieval of the task response. However, theoretical proposals that would allow spreading-activation to decay gradually (e.g., van Maanen & van Rijn, 2007) would not only support the compatibility effect in experiment 4a but also make predictions regarding how long the delay in 4b would have to be before the effect disappeared.

Conclusions

Tucker & Ellis interpret their results through a lens of strict embodiment (e.g., Barsalou, 1999). They argue that the phenomenon could not be due to the perceptual priming of the motor response, rather posit that the evidence supports activation of a more general representation that incorporates both visual and motor properties (Tucker & Ellis, 2001).

Mahon & Caramazza (2008) counter that this line of reasoning unjustifiably discounts the possibility that abstract/symbolic systems could account for visual-motor priming by the spreading activation through abstract

symbols. They propose that the challenge for abstract/symbolic systems is to "1) develop a model of the computations and representations that mediate between perceptual processing and motor activation, and 2) specify the conditions under which those computations are deployed" (Mahon & Carmazza, 2008). In this paper, we present a cognitive model that addresses both of those challenges while remaining within ACT-R's existing architectural constraints. While ACT-R is a traditional abstract/symbolic system, this work moves the architecture towards one that is "grounded by interaction", allowing it to not only exploit the flexibility of disembodied abstractions but also the richly contextualized representations inherent in more strictly embodied accounts (Mahon & Carmazza, 2008).

Combining the generalization of activation spread and co-occurrence associations allows ACT-R to account for semantic (Anderson & Reder, 1999), visual-motor Tucker & Ellis, 2001), and potentially even motor-visual (Craighero, Fadiga, Umilta, & Rizzolatti, 1999) priming. This richer account may also be a fundamental component in enabling symbol acquisition/grounding within ACT-R (Barsalou, 2003; Mahon & Caramazza, 2008).

Acknowledgments

This work was performed while the first author held a National Research Council Research Associateship Award and was partially supported by the Office of Naval Research under job order number N0001408WX30007 and 09-Y861 awarded to the second author. The views and conclusions contained in this document should not be interpreted as necessarily representing official policies, either expressed or implied, of the U.S. Navy.

The models and videos are available for download at *http://www.nrl.navy.mil/aic/iss/aas/CognitiveRobots.php.*

References

Anderson, J. R. (1974). Retrieval of propositional information from long-term memory. *Cognitive Psychology*, 5, 451-474.

Anderson, J. R. (2007) *How Can the Human Mind Occur in the Physical Universe?* New York: Oxford University Press.

Anderson, J. R. & Reder, L. M. (1999). The fan effect: New results and new theories. *Journal of Experimental Psychology: General*, 128, 186-197.

Barsalou, L. W. (1999). Perceptions of perceptual symbols. *Behavioral and brain sciences*, *22*(04), 637-660.

Barsalou, L.W. (2003). Abstraction in perceptual symbol systems. *Philosophical Transactions of the Royal Society of London: Biological Sciences, 358,* 1177-1187.

Breazeal, C. (2009). MDS Robot. http://robotic.media.mit.edu/projects/robots/mds/overview

Craighero, L., Fadiga, L., Rizzolatti, G., & Umilta, C. (1999). Action for perception: a motor-visual attentional effect. *Journal of experimental psychology: Human perception and performance, 25*. 1673-1692.

Fransen, B.R., Herbst, E., Harrison, A.M., Adams, W., Trafton, J.G. (2009) *Real-time face and object tracking*. Proceedings from 2009 IEEE/RSJ international conference on intelligent robots and systems.

Kato, H. Billinghurst, M., Pouplrev, I., Imamoto, K., & Tachibana, K. (2000). Virtual object manipulation on a table-top AR environment. In *IEEE and ACM International symposium on augmented reality*. 111-119.

Mahon, B.Z., & Caramazza, A. (2008) A critical look at the embodied cognition hypothesis and a new proposal for grounding conceptual content. *Journal of physiology - Paris, 102,* 59-70.

Newell, A., & Simon, H. A. (1972). *Human problem solving*. Englewood Cliffs, NJ: Prentice-Hall.

Niedenthal, P. M., Barsalou, L. W., Winkielman, P., Krauth-Gruber, S., & Ric, F. (2005). Embodiment in attitudes, social perception, and emotion. *Personality and Social Psychology Review*, *9*(3).

Pylyshyn, Z. W. (1984). *Computation and cognition*. Mit Press Cambridge, MA.

Shepard, R., & Metzler, J. (1971). Mental rotations of three-dimensional objects. *Science, 171*. 701-703.

Trafton, J.G., Harrison, A.M., Fransen, B.R., & Bugajska, M. (2009) An embodied model of infant gaze-following. In A. Howes, D. Peebles, R. Cooper (Eds.), *9th International conference on cognitive modeling - ICCM2009*, Manchester, UK.

Tucker, M., & Ellis, R. (1998). On the relations between seen objects and components of potential actions. *Journal of Experimental Psychology-Human Perception and Performance*, *24*(3), 830-846.

Tucker, M., & Ellis, R. (2001) The potentiation of grasp types during visual object categorization. *Visual cognition, 8,* 769-800.

Van Maanen, L., & Van Rijn, H. (2007). An accumulator model of semantic interference. *Cognitive Systems Research*, 8(3), 174-181.

Wexler, M., Kosslyn, S.M., & Berthoz, A. (1998). Motor processes in mental rotation. *Cognition, 68,* 77-94.

Wilson, M. (2002). Six views of embodied cognition. *Psychonomic Bulletin & Review*.