Memory Use During Hand-Eye Coordination

Mary M. Hayhoe, Dana H. Ballard, and Steven D. Whitehead
Center for Visual Science
The University of Rochester
Rochester, New York 14627-0226 USA

Abstract[1]

Recent successful robotic models of complex tasks are characterized by use of deictic primitives and frequent access to the sensory input. Such models require only limited memory representations, a well-known characteristic of human cognition. We show, using a sensori-motor copying task, that human performance is also characterized by deictic strategies and limited memory representations. This suggests that the deictic approach is a fruitful one for understanding human brain mechanisms; it also suggests a computational rationale for the limitations on human short term memory.

Models of visual and cognitive processes have traditionally assumed elaborate viewer- or world-based internal models of the environment that are sufficiently rich to allow complex reasoning about the effects of action sequences. However, there are inherent limitations to these models. It is computationally expensive to represent a vast array of time-varying environmental information. Consequently the range of behaviors that can be modeled is very restricted. More recently a number of researchers have demonstrated that a range of complex tasks can be modeled efficiently using very limited memory representations. The key aspect of these models is that complex internal representations are avoided by allowing frequent access to the sensory input during the problem-solving process (Brooks, 1991; 1986; Whitehead & Ballard, 1990; Agre and Chapman, 1987; Ballard, 1989; 1991; Chase & Simon, 1973). The models use "deictic" primitives,[2] which dynamically refer to points in the world with respect to their crucial describing features (e.g., color or shape). The limited memory representations that emerge naturally from these models capture one of the fundamental features of human cognition: the limited nature of short term memory. This has been a focus of research since Miller's seminal article (Miller, 1956). In addition, the saccadic eye movement system provides a natural biological mechanism for efficient access to task-relevant information that is tied to ongoing behavior. Although the comparison between humans and robotic models is suggestive, we have little knowledge of how humans actually perform in tasks comparable to those modeled by robotic systems. We ask here whether human sensori-motor performance can also be characterized by the use of deictic strategies and limited memory representations that has proved so powerful in the formal models. We report evidence from a copying task that reveals that human performance exhibits exactly these characteristics: extremely limited memory representations and the crucial role of eye movements in defining the reference for deictic instructions. The environment can be used as an external store, since frequent access can be made by fixational eye movements. Consequently, complex behaviors can be generated using a small number of simple primitive instructions, without the need for complex reasoning. This suggests that the approach using deictic representations is a fruitful way of understanding human brain mechanisms. It also suggests a computational rationale for why short term memory is limited.

In order to examine the role of eye movements in the performance of complex tasks, we chose a task which involved copying a pattern of colored blocks. The task was chosen to reflect basic sensory, cognitive, and motor operations involved in a wide range of human performance. A display of colored blocks was divided up into three areas, the *model*, *source*, and *workspace*. The model area contains the block configuration to be copied; the source contains the blocks to be used; and the workspace is the area

[1] Supported by AFOSR Grant 91-0332 and NSF Grant IRI-8903582.

[2] The word deictic means "pointing" or "showing" and was first used in this context by Agre and Chapman, building on work by Ullman (1984). It means that aspects of the scene can be referred to by denoting that part of the scene with a special marker. One such marker can be fixation itself: looking directly at a part of the scene provides special access to the features immediate to the fixation point. Thus, an instruction might be to get 'the color of the thing currently fixated' rather than 'the color at location (x,y)'.

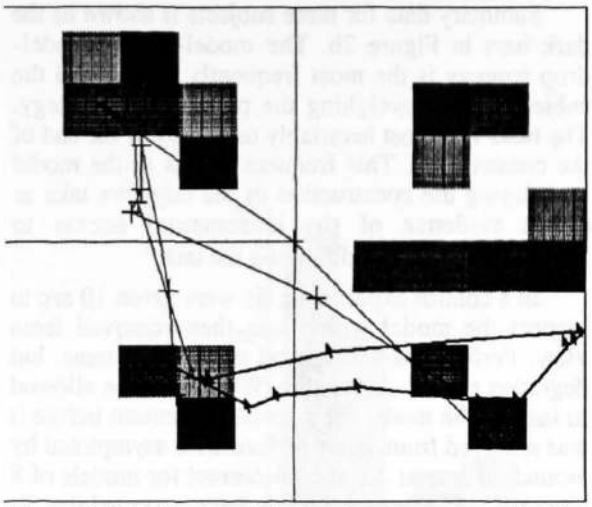

Figure 1. Display used in the hand-eye coordination experiments. The model is displayed on the top left, and the source area is on the right. The bottom left is the workspace for copying the model pattern. The subject's instructions are to build a copy of the model in the workspace area using blocks from the source area. Blocks are moved using a cursor that is controlled by the Macintosh Mouse®. The display subtended 17 by 13 deg of visual angle, individual blocks were 1.7 by 1.3 deg. The right eye is tracked in the experiment. The eye position trace is shown by the cross and the thin line. The cursor trace is shown by the arrow and dark line. Shown is a single cycle, from dropping off block two to dropping off block three (in the experimental trial the blocks appear colored). Immediately after dropping off block two (light gray) the fixation point is transferred to the model, presumably to gain information on the next block. Simultaneously, the cursor is moved to the source area at the extreme right of the screen. Subsequently the fixation point is transferred to the source area at the location of block three (dark gray) and used to direct the hand for a pickup action. Then the eye goes back to the model and the cursor is moved to the drop-off location. The eye moves to the drop-off location to facilitate the release of the block. This display is accomplished by a "replay" program that retraces the experimental course from saved data. In the experiment itself the block is erased immediately after it has been picked up, but for the figure it has been left visible to mark its location. Thus the block moved appears twice: once in the source area and once in the extreme left of the workspace.

where the copy is assembled. Subjects used the cursor driven by a mouse to "pick up" and "place" blocks on the screen. Picking up a block is accomplished by moving the cursor over the block and depressing a button attached to the mouse. Placing the block is accomplished by moving the block to the desired location and releasing the button[3]. Both the eye and cursor movements were monitored throughout the task.[4]

When the task is performed while the model is visible throughout the trial, observations of individual eye movements suggest that information is acquired incrementally during the task and not acquired *in toto* at the beginning of the task. For example, the trace for the third block used by subject K is depicted in Figure 1. Initially the mouse movement and fixation point movements are in different directions, with the cursor being transferred to the source and the eye directed towards the model. The fixation point then moves to the source area at the location of block three (black) and is used to direct the hand for a pickup action. Then the eye *goes back* to the model while the cursor is moved to the workspace. The eye moves to the drop-off location to facilitate the release of the block.

The fact that fixation is used for picking up and dropping off each block would have been expected from data on single hand-eye movements (Milner & Goodale, 1991). However, the extent to which the eyes were used to check the model was unanticipated. It seems likely that humans use their ability to fixate to simplify the task in two ways. First, the "fixation frame" allows the use of deictic primitives. For example, an object is picked up by first looking at it and then directing the hand to the center of the fixation coordinate frame. We call this the *do it where I'm looking* strategy. The alternative requires programming a command in a world- or ego-based coordinate representation, with much greater

---

[3]For this reason the block copying task used a set of coarse-grained, discrete locations for the blocks. Thus releasing the mouse button placed the block at the nearest discrete grid location. This obviated the need for very precise positioning and made the task easier to perform. Block sizes varied from 1/2° to 2°. Using 1.7° by 1.3° blocks, the resultant grid was a 10 by 10 array as can be inferred from Figure 1, which shows the initial configuration for such an example. Displays were random configurations of 8 blocks of four colors: red, green, yellow, blue.

[4]The eye movements were monitored using a Dual-Purkinje Image eye tracker, sampling the eye movements and hand movements every 20 msec. The head was held fixed throughout the experiment, using a bite bar. At the outset of each set of trials for an individual subject, the subject's gaze was calibrated by measuring the recording signal over a grid of 25 positions that spanned the display screen. The accuracy of the tracker is considerably better than one degree so that fixations of individual blocks could be detected with high confidence.

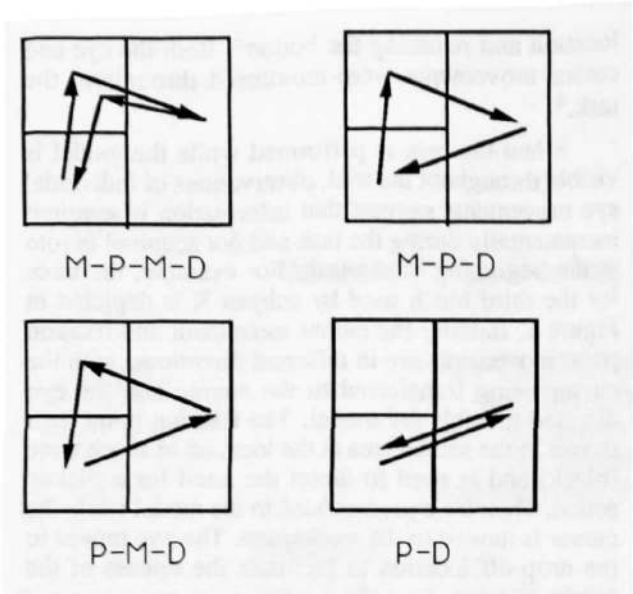

Figure 2a. The different categories of eye movements used in the task. "M" means that the eyes are directed to the model. "P" and "D" mean that the eyes and mouse are coincident at the pickup point and drop-off point respectively. Thus for the PMD strategy, the eye goes directly to the source for pickup, then goes to the model area, and then to the source for drop-off.

demands on the fidelity of the representation. Second, fixation is be used to acquire information *en route* at the point at which it is required. For example, consider the color of the third block. If this is memorized at the outset along with several other colors, then a corresponding number of memory locations would be required. However, a single location that encodes *the-color-of-the-next-block* can be used if the loading of that location is performed at the appropriate moment in the task.

The basic cycle from the point just after a block is dropped off to the point where the next block is dropped off provides a convenient way of breaking up the task into component subtasks of single block moves. This allows us to explore the different sequences of primitive movements made in putting the blocks into place. A way of coding these subtasks is to summarize where the eyes go during a particular subtask. Thus the sequence in Figure 1 can be encoded as "model-pickup-model-drop" (M-P-M-D on the graph legend) with the understanding that the pickup occurs in the source area and the drop in the workspace area. Four principal sequences of eye movements can be identified, as shown in Figure 2a.[5]

Summary data for three subjects is shown as the dark bars in Figure 2b. The model-pickup-model-drop strategy is the most frequently used by all the subjects, far outweighing the pickup-drop strategy. The latter is almost invariably used only at the end of the construction. This frequent access to the model area during the construction of the copy we take as direct evidence of the incremental access to information in the world during the task.[6]

In a control experiment, Ss were given 10 sec to inspect the model which was then removed from view. Performance was good up to four items, but degraded rapidly above this. When Ss were allowed to inspect the model for a variable duration before it was removed from view, performance asymptoted by around 10 secs at about 60% correct for models of 8 blocks. On this basis we might have expected that Ss would have use memory more extensively in the main experiment, but they clearly use only minimal memory when they are free to do so.

In this task the crucial information is the color and relative location of each block. The observed sequences (in the main experiment) can be understood in terms of whether the subject has remembered either the color and/or the location of the block currently needed. The necessary assumptions are that: (1) the information is most conveniently obtained by explicitly fixating the appropriate locations in the model, and (2) the information is preferentially acquired sequentially. If both the color and location are needed, an MPMD sequence should result. If the color is known, a PMD sequence should result; if the location is known, an MPD sequence should result; if both are known, then PD. In the data the PD

[5]It is possible that these frequent eye fixations are an artifact of the use of the computer mouse to move the blocks, which may slow down performance. Perhaps if real three-dimensional blocks were used, the results would be different. We think this unlikely for two reasons. First, even though different Ss used very different hand speeds, they used similar strategies. Second, the task was especially desgned to make block manipulation easy. (In fact, easier than in the three-dimensional case where the fingers take up space and manipulation becomes more difficult.)

[6]This result also points to the use of eye movements as an integral part of the economical execution of the task. What if the subjects had to perform the task while holding their gaze fixed? As a control, we had subjects do this: the model was kept visible but subjects had to fixate the center of the display throughout the task. They were able to complete the task successfully, but required about three times longer. We conjecture that this is not due to difficulty in seeing the blocks (which can be up to 5 degrees eccentric), since we varied the size of the blocks during this control and found that, for sizes in the range of one degree, the time to complete the task is invariant to variations in block size of plus and minus a factor of two.

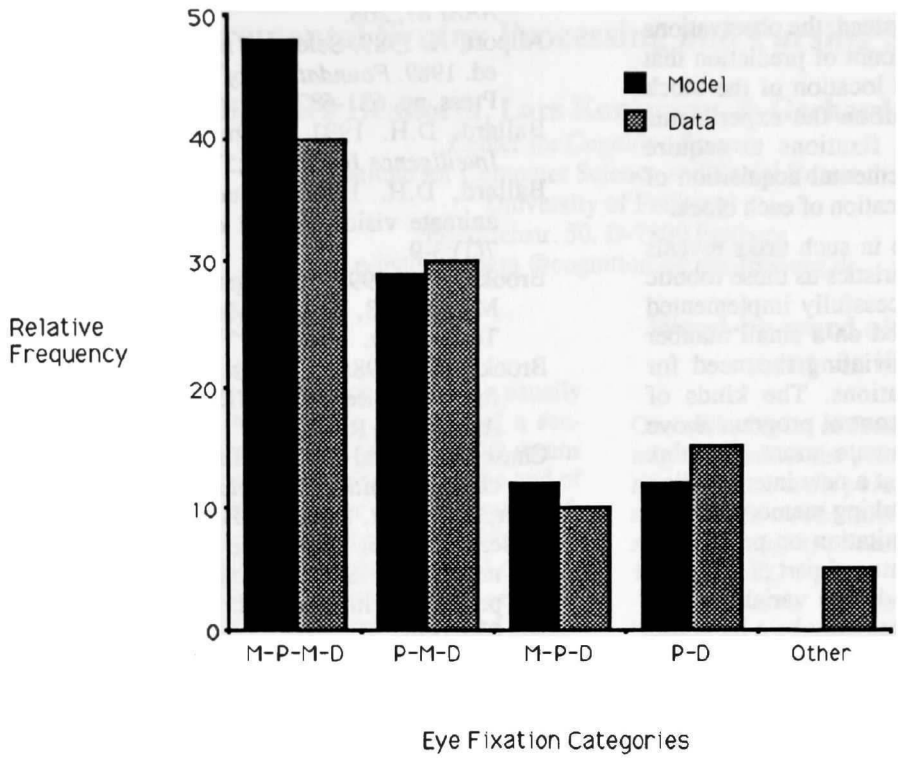

Figure 2b. Frequency of category use for a sample of 100 block moves from 4 observers. Treating the addition of a block to the figure being built in the workspace allows the comparison of different strategies. A strategy that memorized the model configuration at the outset could then consist entirely of pickup and drop operations. Instead, the data summary shows a number of different programs. Comparison of model and data. Model used $P_C = 0.21$ and $P_I=0.11$.

sequences were invariably the last one or sometimes two blocks in the sequence. The MPD and PMD sequences can be explained if the subjects are sometimes able to remember an extra location and/or color when they fixate the model area. To model this effect, we allowed color and location memory to have a capacity of zero to two items. If either of these locations are empty the eyes are drawn to the model area. To explain this in more detail, consider the model control program:

Repeat until {pattern copied}
    GetColor
    PickUp
    GetLocation
    Drop

If each of these basic instructions required fixation, each of the observed sequences would be of the form MPMD. The instructions GetColor and GetLocation act as producers of color and location information respectively. Similarly, PickUp and Drop act as consumers of color and location information. To explain the observed sequences, we only have to allow the model fixations to probabilistically produce an extra color and/or location. The new program becomes:

Repeat until {pattern copied}
    If (no colors in memory) then GetColor
    PickUp
    If (no locations in memory) then GetLocation
    Drop

where now GetColor and GetLocation, in addition to determining a single color or location, are each allowed to determine the subsequent color with probability P_C and the subsequent location with probability P_I. The only remaining modification is that at the penultimate block, subjects invariably memorize the color and location of the last block. When this feature is added to the model as a special case the observed data can be modeled very closely, as shown by the gray bars in Figure 2b.

In summary, the main result is that the information required for the task is acquired just prior to its use. The alternate strategy of memorizing the configuration to be copied in its entirety before moving blocks is never used. It is never used even though it technically could be: our memory experiments show that up to four blocks can be copied from memory without error. Instead, the observations point to the use of a small amount of prediction that only extends to the color and location of the block after the current one. In addition the experiments point to (1) the use of eye fixations to acquire information, and (2) the incremental acquisition of information of the color and location of each block.

Thus human performance in such tasks reveals the same fundamental characteristics as those robotic models which have been successfully implemented using deictic instructions based on a small number primitive operations, thus obviating the need for complex memory representations. The kinds of primitives used in the simple control program above can clearly be used to generate more complex behaviors. These results suggest a new interpretation of the limitations of human working memory. Rather than being thought of as a limitation on processing capacity, it can be seen as an integral part of a system which makes dynamic use of deictic variables. The limited number of variables need only be a handicap if the entire task is to be completed from memory; in that case the short term memory system is overburdened. In the more natural case of performing the task with ongoing access to the visual world, the task is completed perfectly. This suggests that a natural metric for evaluating behavioral programs can be based on their spatio-temporal information requirements.

These results also suggest a new interpretation of the role of foveating eye movements in vision. Rather than being thought of as a consequence of the poor resolution of peripheral vision, fixation can be seen as defining the variable currently relevant for task performance and thus orchestrating performance of the task. Similarly, the attention can be seen as being necessarily limited by virtue of its role in specifying the variable for the next instruction.[7]

Historically we have been accustomed to thinking of the job of perception as creating rich, task-independent descriptions of the world which are then re-accessed by cognition (Marr, 1982). However, an intriguing suggestion that follows from these experiments is that perhaps the job of perception can be greatly simplified: it need only create descriptions that are relevant to the current task (see also (Nakayama, 1990)).

[7]A similar suggestion has been made by (Allport, 1989).

REFERENCES

Agre, P.E. and Chapman, D. 1987. Pengi: An implementation of a theory of activity. In *Proc., AAAI* 87, 268.

Allport, A. 1989. Selective attention. In Posner, M., ed. 1989. *Foundations of Cognitive Science.* MIT Press, pp. 631-682.

Ballard, D.H. 1991. Animate vision. *Artificial Intelligence Journal* 48:57-86.

Ballard, D.H. 1989. Behavioral constraints on animate vision. *Image and Vision Computing* 7(1):3-9.

Brooks, R.A. 1991. Intelligence without reason, AI Memo 1293, AI Lab, Massachusetts Institute of Technology.

Brooks, R.A. 1986. A robust layered control system for a mobile robot. *IEEE J. Robotics and Automation* RA-2:14.

Chase, W.G. and Simon, H.A. 1973. Perception in chess. *Cognitive Psychology* 4:55-81.

Marr, D.C. 1982. *Vision.* W.H. Freeman and Co.

Miller, G. 1956. The magic number seven plus or minus two: Some limits on your capacity for processing information. *Psychological Review* 63:81-96.

Milner, A.D. and Goodale, M.A. 1991. Visual pathways to perception and action. Ctr. for Cognitive Science, U. Western Ontario, COGMEM 62.

Nakayama, K. 1990. The iconic bottleneck and the tenuous link between early visual processing and perception. In Blakemore, C., ed. *Vision: Coding and Efficiency.* Cambridge Univ. Press, pp. 411-422.

Ullman, S. 1984. Visual routines. *Cognition* 18:97-157; also in Pinker, S., ed. 1984. *Visual Cognition.* Cambridge, MA: Bradford Books.

Whitehead, S.D. and Ballard, D.H. 1990. Active perception and reinforcement learning. *Neural Computation* 2(4):409.