PHILOSOPHICAL TRANSACTIONS OF THE ROYAL SOCIETY B BIOLOGICAL SCIENCES

Hand-Eye Coordination during Sequential Tasks [and Discussion]

Dana H. Ballard, Mary M. Hayhoe, Feng Li, Steven D. Whitehead, J. P. Frisby, J. G. Taylor and R. B. Fisher

Phil. Trans. R. Soc. Lond. B 1992 **337**, 331-339
doi: 10.1098/rstb.1992.0111

Hand–eye coordination during sequential tasks

DANA H. BALLARD[1], MARY M. HAYHOE[2], FENG LI[3] AND STEVEN D. WHITEHEAD[1]

[1] *Computer Science Department, University of Rochester, Rochester, New York 14627-0226, U.S.A.*
[2] *Center for Visual Studies, University of Rochester, Rochester, New York 14627-0226, U.S.A.*

SUMMARY

The small angle subtended by the human fovea places a premium on the ability to quickly and accurately direct the gaze to targets of interest. Thus the resultant saccadic eye fixations are a very instructive behaviour, revealing much about the underlying cognitive mechanisms that guide them. Of particular interest are the eye fixations used in hand–eye coordination. Such coordination has been extensively studied for single movements from a source location to a target location. In contrast, we have studied multiple fixations where the sources and targets are a function of a task and chosen dynamically by the subject according to task requirements. The task chosen is a copying task: subjects must copy a figure made up of contiguous coloured blocks as fast as possible. The main observation is that although eye fixations are used for the terminal phase of hand movements, they are used for other tasks before and after that phase. The analysis of the spatial and temporal details of these fixations suggests that the underlying decision process that moves the eyes leaves key decisions until just before they are required.

1. INTRODUCTION

Robotics researchers have long been interested in complex sensorimotor tasks, but have had great difficulty in constructing systems to perform or learn to perform such tasks. It has often been proposed that complex behaviours should be achieved by combining more primitive ones, but this observation has been insufficient to solve the problem. One complication has been that primitives are often designed to produce and use internal models that have proven inadequate to handle the dynamic nature of the world. The basic strategy has been to focus on the reasoning component of the problem solving and neglect the sensory-motor aspect. To make this strategy work, it is assumed that an internal model can be constructed that is sufficiently complex to project the effects of sequences of actions into the future. This research program has had considerable success in problem solving environments that require minimal interaction with the world, such as chess, but much less success in very dynamic situations. The principal difficulties are the basic computational complexity of the reasoning component and in the adequacy of formal logical models to capture the current state of the world as well as the range of possible futures.

More recently a number of researchers (Brooks 1991; Ballard 1991; Bajcsy 1988) have argued that a significant part of the behavioural repertoire can be modeled by allowing frequent access to sensory input during the problem solving process. In this way the computational complexity of the world can be avoided by the use of primitives that dynamically reference the structure of the world. Such primitives have been termed *deictic*. Using this strategy, sequences of primitives can succinctly create complex behaviours as each primitive implicitly defines the context for its successor. These strategies are particularly useful for behaviours that are repeated, as the decision making process can learn from previous situations.

Part of the motivation for the study of sequential behavioural strategies has come from the structure and function of human vision. The human eye is distinguished from current commercial electronic cameras by virtue of having much better resolution near the optical axis. It has a high-resolution fovea where over a one-degree range the resolution is better by an order of magnitude than that in the periphery. One feature of this design is the simultaneous representation of a large field of view and high acuity in the fovea. With the small fovea at a premium in a large visual field, the human visual system has special fast mechanisms (saccades) for moving the fovea to different spatial targets. Most of the previous work in computational vision has avoided this complexity, with the result that although we have rich models for the computation that could be occurring during a single fixation (Marr 1982), much less is understood about how vision interacts with cognitive decision making. In this latter context, it is instructive to examine the structure and function of saccadic eye movements in the process of solving complex tasks. The first systematic study of saccadic eye movements in the context of behaviour was done by Yarbus (1967). When subjects were given the task of remembering the position of the people and the objects in the room the eye movement traces showed a specialized signature for this task that was not similar to other signatures, say for the one elicited for 'give the ages of the people in the picture'. Because of this, we conjecture that representing large amounts of categorical data about our visual world is not done routinely. Instead, the visual system may be used to subserve problem-solving behaviours that may or may not require an elaborate model of the world in the traditional sense of remembering positions of people and objects in a room.

This view is supported by a large body of data concerning human short term memory (Baddeley 1986). Under a variety of different experimental conditions it has been demonstrated that subjects have difficulty remembering more than about seven novel items. Viewed as simply a memory task, this seems limiting; however, in this paper we offer a different interpretation of these data. Our principal hypothesis is the following: if the memory items are seen as variables in a behavioural repertoire, then the number of behaviours scales exponentially with the number of items. Thus a small number of variables can be used to produce an extremely large number of behaviours. The costs associated with discovering good behaviours in this set scale with the number of variables, leading to a premium on keeping the number of variables small.

We have studied our principal hypothesis in the context of hand–eye coordination using the following experiment. A display of coloured blocks is 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 where the copy is assembled. Subjects use 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.

Because humans can fixate an environmental point, they can use a special frame of reference centred at that point to simplify this task. The 'fixation frame' is viewer-oriented, but not viewer-centred. (The origin of this frame is at the point of intersection of the two optical axes. To orient this frame, one axis can be parallel to the line joining the two camera centers and another can be chosen as the optical axis of the dominant eye.) It allows for closed loop behavioural strategies that do not require very precise three-dimensional information. For example, an object can be picked up by first looking at it and then directing the hand to the centre of the fixation coordinate frame. Informally, we refer to this behaviour as a 'do-it-where-I'm-looking' strategy, but more technically it is referred to as a deictic strategy after Agre & Chapman (1987).

The deictic strategy of using the perceptual system to actively control the point of action in the world has precisely the right kind of invariance for a large number of behaviours. As an example, consider the block copying problem in more detail. To copy the first block, its colour has to be determined. One way to do that is to fixate the block in question in the model area. Next a block of that colour has to be located and picked up in the source area. Finally that block has to be moved and dropped in the workspace. Subsequent block moves are more complicated as the relationship of the new block with the model has to be determined and then replicated in the workspace. These functions perhaps could be all done in a single step but the sequential nature of the hand movements argues that the simplest model will exploit this

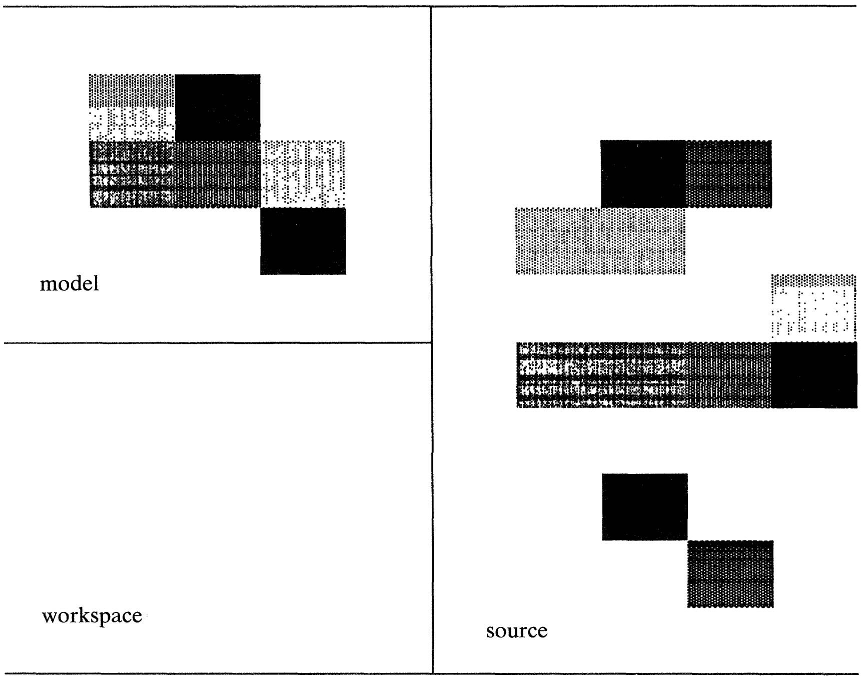

Figure 1. Display used in the hand-eye coordination experiments. The subject's instructions are to build a copy of the model in the workspace area using blocks from the source area. Blocks are moved using a cursor that is controlled by the Macintosh Mouse^{TM}.

sequentiality. So in rough outline a cognitive program for a block move could look like:

Repeat until {the pattern has been copied}:

  Fixate (*A block in the model area with a given colour*)
  Remember (*Its location within the context*)
  Fixate (A block in the source area with the same colour)
  Pickup (The block currently fixated)
  Fixate (The appropriate location in the workpace)
  Move (The block to the fixated location)
  Drop (The block currently held at the current location)

In the program italics are used to denote a structure that has to be 'bound' to the sensor data, whereas non-italicized arguments to the primitives denote structures that are already bound. Here binding is used in the same way as the value of a variable in a computer program is bound or assigned to that variable. These instructions make extensive use of deictic reference. It is assumed that the instruction *Fixate* will orient the center of gaze to point to a place in the image with a feature appropriate to the argument of the function. If multiple instances of the feature are present, then some tie-breaking scheme must be used. The deictic nature of these instructions is illustrated by *Pickup*, *Move*, and *Putdown*, which are assumed to act at the centre of the fixation frame, irrespective of its specific location in three-dimensional space. *Fixate* is actually also a deictic instruction if we include a mechanism for target selection, such as a focus of attention. A focus of attention may be thought of as an electronic fovea, in terms of its ability to select target locations. Thus *Fixate* becomes two instructions:

  Attend To (*image-feature appropriate for target selection*)
  Fixate (Attended location)

The exact behaviour of these instructions depends on the particular embodiment, but the assumption at this level of modeling is that they can be designed to perform as described. We have shown that similar albeit simpler problems can be learned by computer using these instructions (Whitehead & Ballard 1990).

The simplifications of deictic strategies may be understood in terms of computational complexity. Thinking generally about the problem of relating internal models to objects in the world, one way to interpret the need for sequential, problem-dependent eye movements is as a suggestion that the general problem of associating many models to many parts of the image simultaneously is too hard. To make it computationally tractable within a single fixation, it has to be simplified, either into a problem of location (one internal model) or identification (one world object). Table 1 summarizes this view and also classifies our deictic primitives. This table suggests that a visual task can be solved by partitioning the task into two different kinds of subtasks. We model the visual organization more crudely into a centre and a surround. The centre is 'where-I'm-looking' and the surround is a source of new gaze points. A location task is to find the image coordinates of a single model in the presence of many alternatives. In this task the image periphery must be searched. One can assume that the model has been chosen *a priori*. An identification task is to associate the foveated part of the image with one of many possible models. In this task one can assume that the location of the material to be established is at the fixation point. Work using colour cues has shown that this simplification leads to dramatically faster algorithms for each of the specialized tasks (Swain & Ballard 1991). Thus we think of the eye movements as solving a succession of location, manipulation, and identification subtasks in the process of meeting some larger cognitive goal.

Table 1. The organization of visual computation into WHAT/WHERE modules may have a basis in complexity. Trying to match a large number of image segments to a large number of models at once may be too difficult to accomplish within a single fixation

image parts	models	
	one	many
one	manipulation: trying to do something with an object whose identity and location	identification: extracting properties of an object whose location can be fixated
	Pickup, Move, Drop	Remember
many	location: trying to find a known object in a wide field of view	too difficult to accomplish within a single fixation?
	Move, Fixate, Attend To	

2. BACKGROUND

The study of complete tasks is a necessity in robotics owing to the task demand of synthesizing a complete system, but has less of a tradition in psychology. Many experimental paradigms have resorted to reduced stimuli and task configurations in an attempt to reduce the number of possible hypotheses. However, one danger of this strategy is that subjects' behaviour is often obtained under very unnatural conditions. Studying the human system in extremis may not generate results that extend to natural conditions. This is the main argument for using natural experimental protocols and trying to deal with the resultant rich hypothesis space. Such protocols have been used in the study of reading and inspired the copying task used in this study.

Short term memory was originally conceptualized as a limited capacity buffer or store which could hold a small number of items. A more useful view emerged in the 1970s (Baddeley & Hitch 1974) which emphasized the functions served by short term memory. Short term memory is now commonly viewed as reflecting the cognitive and perceptual operations the subject is currently involved in. That is, the fundamental limitations appear to be in the operations or processes that one can perform, as well as in the perceptual information used by these processes (e.g. Baddeley's visuo-spatial 'scratch pad' (Baddeley 1986)). However, very little is known about the nature of these central processing limitations. In this paper we view the limitations as being set by the number of variables that can easily be bound to programs. The attraction of the copying task is that it segments naturally into subtasks involving individual blocks, and this segmentation must be dealt with directly by the subject as only one block can be moved at a time. This provides a constrained background for guessing the cognitive state of the subject relative to the task.

The use of deictic strategies owes its recent history to a novel object-centred processing strategy introduced by Ullman (1984) termed 'visual routines'. This was greatly extended by Agre & Chapman (1987), who developed much of the rationale for its use. The strategy uses only a few object-centred variables called 'markers' and defines a library of primitive operations on them. Subsequently, we have shown that in the context of a simulated blocks world, difficult tasks that require vision can be successfully *learned* (Whitehead & Ballard 1990, 1991). The key idea is that by allowing the system the freedom to select the key task-relevant features, succinct descriptions of the task can be created which have transfer to novel situations.

3. EXPERIMENTAL DESIGN

The main point of our experimental design was to be able to study hand–eye coordination over protracted periods where subjects made multiple volitional movements. We wanted a task where subjects were free to choose appropriate eye and hand movements in a way that was constrained only by the task requirements. Much previous work has been done on the case where a subject makes a single eye and/or hand movement to a target often accompanied by contingent instructions (Jeannerod & Decety 1990).

We were anxious that the task focus on decision making and not fine motor coordination. For this reason the block copying task used a set of coarse-grained, discrete locations for the blocks. Thus when the mouse button was released, the block would be placed 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° blocks, the resultant grid was a 10 × 10 array as can be inferred from figure 1, which shows the initial configuration for such an example.

The eye movements were monitored using a Dual-Purkinjee Image eye tracker, sampling the eye movements and hand movements every 20 ms. 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.

4. RESULTS

Most remarkable is the interleaving of (i) the use of the fixation point to guide terminal mouse movements with (ii) the use of the fixation point to gather information about the copying details. The data set allows the detailed examination of the movements related to a particular block. In fact we have constructed a 'decompiler' program that creates a symbolic description of the output. A detailed examination of these records shows that the computations for each block tend to be kept separate. For example, a block tends not to be fixated until just prior to its use. This allows the examination of the computations between the time just after a block has been dropped off until the subsequent block is dropped off. For example, the trace for the third block used by subject K is depicted in figure 2.

Figure 2 also shows an interesting feature of hand and eye targeting. Initially the mouse movement and fixation point movements are in different directions, with the former being transferred to the source and the latter directed towards the model. This implies that both the mouse movement and eye movement had to have been programmed by some attentional mechanism without using the fixation point to select the target. 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 (black) 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.

Our experimental procedure allows us to analyse the experimental strategies used by different subjects in some detail. These detailed analyses show that subjects consistently refer back to the model throughout the execution of the task in regular ways. For example, figure 3 shows representative data for subject K performing the six-block copying task depicted in figures 1 and 2. The task was performed in approximately 22 s. From these traces, by examining the times when the eyes and hand were coincident, it can readily be seen that the hands and eyes are coincident just before picking up or dropping off a block. The light grey region indicates the times when the eyes are in the model area; the grey region indicates the times when the eyes are in the source area; and the dark grey region indicates the times when the eyes are in the workspace area. The fact that the light grey regions are interleaved with the other regions indicates that the model is referred to throughout the task. This contrasts with a strategy of memorizing the model at the outset. If the latter were used, then the eye movements would presumably lead the hand movements back and forth from the source to the workspace area. 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.

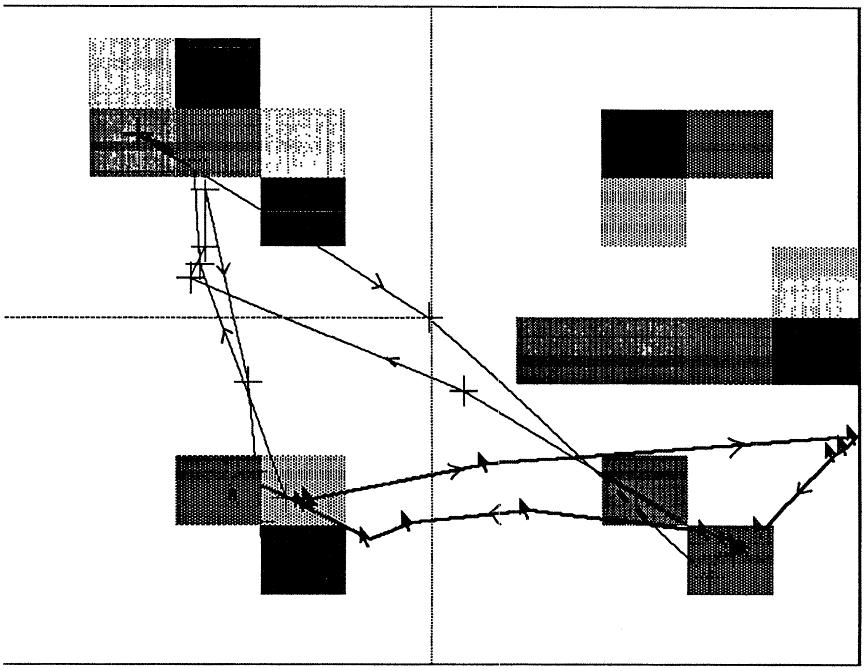

Figure 2. Display used in the hand–eye coordination experiments. The subject's instructions are to build a copy of the model in the workspace area using blocks from the source area. Blocks are moved using the cursor (arrow) that is controlled by the Macintosh Mouse^{TM}. The right eye is tracked in the experiment and its position indicated by a cross on the screen. Shown is a single cycle, from dropping off block two to dropping off block three (in the experimental trial the blocks appear coloured). Immediately after dropping off block two (light grey) 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 grey) 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 area.

Eye velocity exhibits the characteristic high speed profile of saccades. The traces show that the pursuit eye movement system is never used, even though it presumably could have been used, for example to track the progress of the hand movement. The hand velocity exhibits a marked two-phase profile: a ballistic phase, whereby the hand takes off at a high speed toward a target, and a terminal phase where the hand is adjusted prior to a pick-up or drop. The terminal hand velocity phase is indicated by the smaller velocity profile that immediately follows the ballistic profile. The fact that the terminal phase coincides with the period that the hand and eyes are coincident would suggest that this is its purpose.

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. From the point of the deictic representations, we are particularly interested in the amount of memory used for the task. For example, if subjects memorized the model configuration, and then built it with a sequence of pickups and drops between the source area and the workspace area, this would require sufficient memory to remember the entire configuration at once. Of six subjects tested, none used this strategy.

A way of coding these subtasks is to summarize where the eyes go during a particular subtask. Thus the sequence in figure 4 can be encoded as 'model-pickup-model-drop' (or M-P-M-D on the graph legend) with the understanding that the pickup occurs in the source area and the drop occurs in the workspace area. This code can be read by looking at the strip of coarse eye locations in the region appropriate for the third block in figure 3. The summary data for three subjects is shown in figure 4. This shows that 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 only used at the end of the construction. This frequent access to the model during the construction of the copy we take as direct evidence of the incremental access to information in the world during the task.

Further evidence that subjects are encoding features about the immediate task just before the task comes from a detailed inspection of the saccade targets. For example, one can count the number of times that a block that was picked up in the source area was directly preceded by a fixation of a block of the same colour in the model area. This happens for approximately 56% of the occurrences. Another measure is

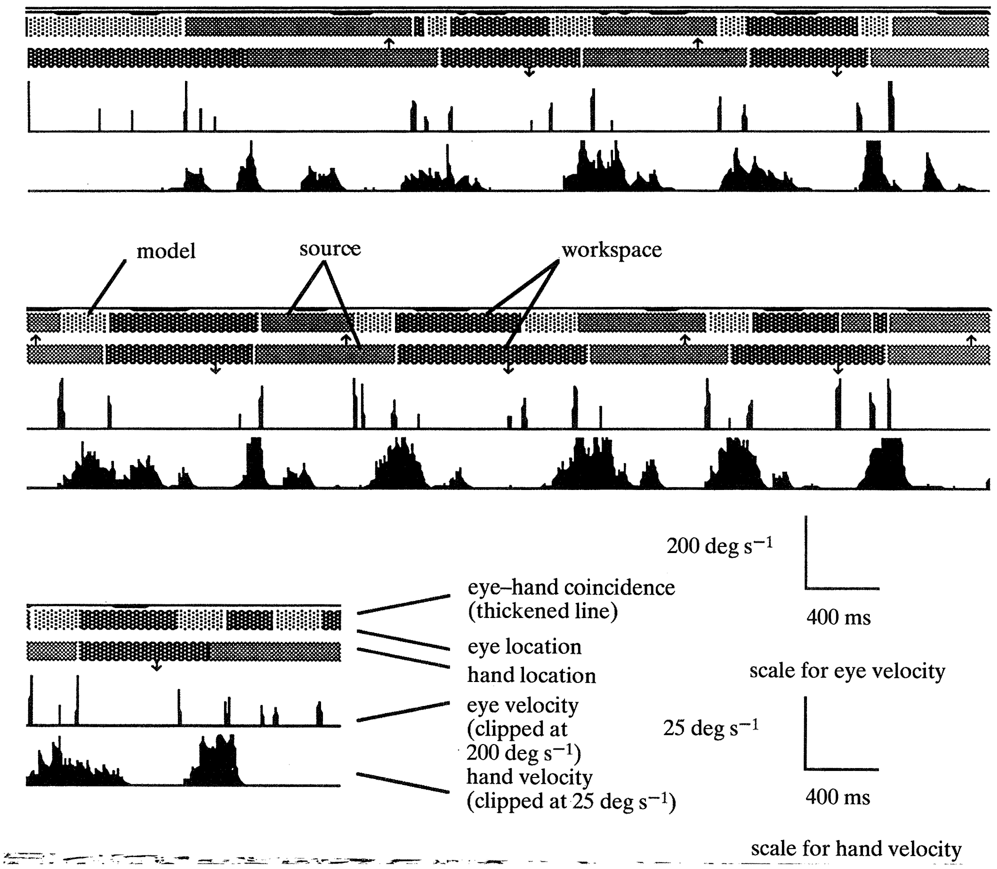

Figure 3. Results for subject K copying the block configuration shown in figure 2. The figure shows three rows of recorded data comprising 22 s of elapsed time. The top single line becomes a double line wherever the eyes and hand are coincident. The two bands just below encode the location of the eyes and hand respectively. Light grey = model; grey = source; dark grey = workspace. For the latter band, an up arrow marks the point where a block has been picked up and a down arrow marks the point where a block is dropped. The next two traces encode eye and hand velocity respectively.

the fixation of the drop-off point. Do the eyes saccade directly to the right point, or do they move instead to a place on the model in the workspace and then use a second saccade to get to the drop-off point? The data shows that on about 30% of the cases, subjects are able to saccade directly to the drop-off point. It could be that the saccades are inaccurate and that the direct fixations of the drop-off point are just part of this inaccuracy, but the fact that the system is capable of recognizing that it has reached the correct point, and does not make a corrective saccade, suggests the use of a model.

All the above data point to the use of a model in the execution of this task. But the central hypothesis is that this model is built up incrementally during the task and not acquired *in toto* at the beginning of the task. To test the latter possibility, we had subjects perform a control task where they were instructed to complete the task from memory. Subjects were given 10 s to look at the model before it was removed from view. In the initial trials of five different patterns of six to nine blocks, none of three subjects was able to complete all five copies correctly. One subject was able to copy three of the patterns correctly, but the other two made errors in every pattern. In an effort to quantify this further, the number of errors as a function of the number of blocks was studied in more detail. Figure 5 shows the results of three trials at each block size for each of four subjects: performance degrades rapidly above four items.

The above model 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 centre of the display throughout the task. They were able to complete the task successfully, but required about three times as much time to complete the task. We conjecture that this is not due to difficulty in seeing the blocks (which can be up to 5 degrees eccentric), as 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. One would like to use this control to circumscribe the task instructions in some way. Unfortunately, it seems that the extra time is required to implement a gaze holding strategy, so that this control does not shed much light on the block copying task per se.

5. DISCUSSION

The principal motivation for the paper was to attempt

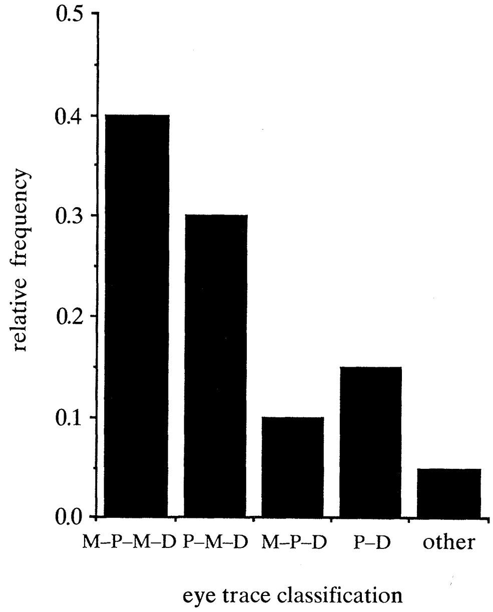

Figure 4. Treating the addition of a block to the figure being built in the workspace allows the comparison of different strategies. A strategy that memorized the model configuration at the outset could then consist entirely of pickup and drop operations. Instead, the data summary, for nearly one hundred block movements, shows a number of different programs. '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.

to analyse the block copying task in terms of a primitive instruction set. In light of the experimental data, let us analyse the program introduced in Section 1. One of the first points of revision concerns the exact timing of the instructions. We assumed that the Drop instruction was done before the next Fixate. In practice this is not the case. The inspection of the record in figure 3 shows that the drop-off physically occurs after the eyes have gone on to the model area. However, it may well be that the decision to drop is made at the time the eyes and hand are coincident, which is shown on the figure to precede the releasing of the button in almost all cases. Thus although the cognitive program may have the sequentiality we hypothesize, the delays in implementing these decisions by the motor program can confound the issue.

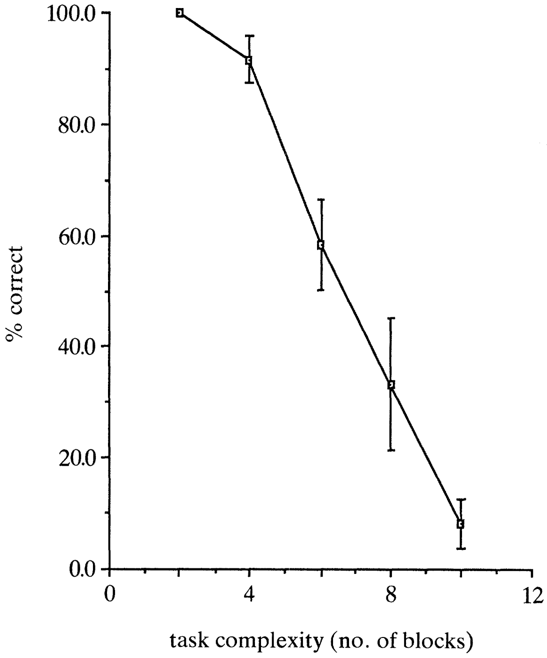

Figure 5. The degradation in copying performance in the memory task as a function of the number of blocks to be copied. Subjects were given 10 s to memorize the configuration before it was removed from view. Error bars show standard errors of the mean value.

Another revision concerns the exact point of 'Remembering'. The initial program used one step to obtain all the relational information, but the experimental data hint that it may be distributed over instances of model fixations. For example our model-pickup-model-drop sequence is more likely encoded as:

```
Repeat until {the pattern has been copied}:
    Fixate (A block in the model area)
    Remember (Its colour)
    Fixate (A block in the source area with the same
      colour)
    Pickup (The block currently fixated)
    Fixate (The block in the model area with the same
      colour)
    Remember (Its relationship with the context)
    Move (The block to the fixated location)
    Drop (The block currently held at the current
      location)
```

The control of copying the model from memory produced an unexpected dividend: During the 10 s exposure phase, the subjects' use of eye movements was an obvious cue to their memorization strategy. They would fixate blocks in rehearsed sequences throughout the period. This suggests that the use of fixation is an aid to the binding process whereby the memory items are affixed in the temporary store. This aspect of the experiment needs to be pursued in more detail.

Our modeling effort has placed a large premium on the success of *AttendTo* which identifies targets for both hand and eye movements. One concern is that the problem of implementing this instruction might be a very difficult one to solve given the general difficulty of computational vision. However, it is possible that this might not be the case if perspicuous image features are used in a novel way. We have used colour in a colour histogram (Swain 1990; Swain & Ballard 1990) to identify image locations in a way that could be a model for an attentional process. In brief, the algorithm uses feedback from the non-retinotopic colour histogram to the retinotopic representations of colour in the following way. The image colours are rated as to how helpful they are in selecting the current model. Local clusters of helpful colours can be found by low-pass filtering and peak detection. This algorithm has been shown to be very effective on tests of databases of multicoloured objects.

6. CONCLUSIONS

Work done with a particular exocentric model of the use of gaze, combined with the concept of deictic representations, suggested that it would be efficacious if the human visual system solved problems in a minimal way. To explore this possibility, hand–eye coordination was studied for a sequential copying task. The strategy of memorizing the configuration to be copied in its entirety before moving blocks seemed to never be used. Instead, a variety of different programs were used to check the course of the copying task in the midst of moving individual blocks. These strategies point to the use of minimal memory solutions. In addition, on a significant number of the trials the hand-directed mouse movement was interleaved so as to be initiated without a prior saccade to the target. This implies that some attentional mechanism can be used for this task, consistent with earlier studies of single movements.

The main observation is that although eye movements are used for the terminal phase of hand movements, they are used for other tasks prior to that phase. This implies that the underlying decision process that moves the eyes leaves key decisions until just before they are required. The results are compatible with the standard model of the brain as having a limited capacity short term memory, but they suggest a new interpretation of that memory. Rather than focus on the small number of items themselves, one should think of the number of behavioural programs that can be constructed with a limited number of memory registers. As this number grows exponentially with the number of registers, there is an advantage to keeping the number of registers small: it allows the systematic discovery of good programs by a search strategy.

This research was supported in part by the National Institutes of Health grant EY01319, the Air Force Office of Scientific Research grant 91-0332, the National Science Foundation grant IRI-8903582, and the Human Science Frontiers Program.

REFERENCES

Agre, P.E. & Chapman, D. 1987 Pengi: An implementation of a theory of activity. *Proc. AAAI* **87**, 268–272.

Baddeley, A. 1986 *Working memory*. Oxford Clarendon Press.

Baddeley, A. & Hitch, G. 1974 Working Memory. In *Advances in learning and motivation*, vol. 8 (ed. G. Bower), pp. 47–90. New York: Academic Press.

Bajcsy, R. 1988 Active perception. *Proc. IEEE* **76**, 996–1005.

Ballard, D.H. 1991 Animate vision. *Artif. Intell. J.* **48**, 57–86.

Brooks, R.A. 1991 Intelligence without reason. AI Memo 1293, AI Lab., MIT.

Jeannerod, M. & Decety, J. 1990 The accuracy of visuomotor transformation: An investigation into the mechanisms of visual recognition of objects. In *Vision and action: the control of grasping* (ed. M. A. Goodale). Norwood, New Jersey: Ablex Pub. Corp.

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

Milner, A.D. & Goodale, M.A. 1991 Visual pathways to perception and action. COGMEM 62, Center for Cognitive Science, University of Western Ontario.

Swain, M.J. 1990 Color indexing. TR 360 and Ph.D. thesis, University of Rochester.

Swain, M.J. & Ballard, D.H. 1991 Color indexing. *Int. J. Comput. Vis.* **7** (1) (Special Issue), 11–32.

Swain, M.J. & Ballard, D.H. 1990 Indexing via color histograms. *Proc. Int. Conf. on Computer Vision* (*ICCV 90*), Kyoto, Japan.

Ullman, S. 1984 Visual routines. *Cognition* **18**, 97–157. (Also in *Visual cognition* (ed. S. Pinker), pp. 97–160. Cambridge, Massachusetts: Bradford Books.)

Whitehead, S.D. & Ballard, D.H. 1990 Active perception and reinforcement learning. *Neural Comput.* **2** (4), 409–419.

Whitehead, S.D. & Ballard, D.H. 1991 Learning to perceive and act by trial and error. *Mach. Learn.* **7** (1), 45–83.

Yarbus, A.L. 1967 *Eye movements and vision*. Plenum Press.

Discussion

J. P. Frisby (*AI Vision Research Unit*, *University of Sheffield*, *U.K.*). Professor Ballard's task of creating patterns of coloured patches on a screen by 'picking up' patches with a mouse has the advantage of simplicity, and hence ease of analysis. But might the arbitrary nature of those pattern configurations, as well as perhaps their two-dimensional character, limit the generality of his results? To what commonplace real-life tasks do you think your experimental paradigm best relates?

D. H. Ballard. Our intent was that the experimental setup would be sufficiently general to be representative of a wide variety of hand–eye tasks. However, the next step would be to prove this by manipulating real blocks.

The mouse-guided movements seem somewhat unnatural, but are they really? Humans have to use tools regularly and this requires making the adjustment of working with a remote coordinate system. And people are good at video games. Our task is also simplified by using a 10 × 10 grid of positions. This means that the patches snap into position without the requirement of much fine motor coordination. Also there is no gravity, slipping, etc. So in some ways this is an easy task. Even so the eyes were used for every pickup and drop.

The most unexpected finding of the experiment was the frequent checking of the pattern be copied, about $1\frac{1}{2}$ times per block. If this is done, as we suspect, to incrementally acquire information about what to do next, this is very likely to be a feature of the more natural task setting.

J. G. Taylor (King's College Centre for Neural Networks, London, U.K.). Can parallel search versus serial search tasks help one to tease out more fully the manner in which combinatorial explosion causes the need for attention?

D. H. Ballard. This is a nice suggestion. Presumably one could redesign the blocks task in a way that some blocks popped out and others did not and find out if there were any revealing differences in search protocol. This may be difficult however due to the scale of overt (eye movements) and covert (serial search independent of eye movements) attention. Overt attention operates on a 200 ms scale and covert attention perhaps at a 20 ms timescale, so the amount of such attention used may be difficult to determine.

R. B. Fisher (*Department of Artificial Intelligence, University of Edinburgh, U.K.*). In Professor Ballard's demonstration of visual search using a colour space, is he not throwing away useful geometric information? That is, a pile of rags with the same colour distribution would produce the same results, but clearly we would not be fooled.

D. H. Ballard. This question refers to our model of object location based on colour indicies (Swain & Ballard 1991). The question touches on many central issues. Perhaps the most central is: what algorithms do humans actually use in vision? The colour algorithm would be fooled by variations in geometry but on the other hand a geometrical algorithm would be 'fooled' in the supermarket, looking at identical rectangular boxes distinguished by different colours. It seems likely that humans can tailor feature selection to the immediate perceptual problem. Colour features are some of the most robust, as they have a large amount of view invariance, a property not shared by geometrical features.

Another question concerns the timecourse of perception. Our immediate concern is to have a fast algorithm that can direct eye movements. This must be done in just a few iterations. Questions about geometry may be asked on much longer timescales. It is not obvious that humans perceive geometry with a lot of fidelity.