A Computer Model of 2D Visual Attention*

Mark Wiesmeyer and John Laird
Artificial Intelligence Laboratory
The University of Michigan

In this paper we present a model of human visual attention that is an extension of a preexisting cognitive theory, the Model Human Processor (MHP) [Card *et al.*, 1983]. The type of visual attention that we model is independent of eye movements and serves as a form of stimulus selection within the visual field. Our goal is to provide a finer grain of reaction time prediction for visual tasks requiring attention than is provided by the MHP. We have developed algorithms based on our model to account for response times of object identification experiments of Colegate, Hoffman and Eriksen (1973) (hereafter CHE). We have implemented these algorithms in Soar [Laird *et al.*, 1987; Laird *et al.*, 1990], which is a reification of some of the basic principles of the MHP, and has been proposed as a unified theory of cognition [Newell, 1990]. Our system models the sequencing of deliberate controllable visual acts of cognition that take on the order of 50 msec. The results of our work suggest that a variant of the "Zoom Lens Model" [Eriksen and Yeh, 1985; Eriksen and St. James, 1986] of visual attention coupled with Soar's theory of deliberate behavior is sufficient for modeling these phenomena, thus, extending the predictive power of the Model Human Processor.

1 The Model Human Processor

The MHP is a cognitive model of human behavior that has been most successfully applied to modeling the timing of man and machine interactions [John, 1988; John and Newell, 1989]. The goal of the MHP is to serve as an engineering tool for estimating human performance. Figure 1 shows a simplified representation of the MHP from Card et al. (1983). On the top left-hand side of the figure is the Perceptual Processor or Perception, which consists of sensory modalities, such as vision and audition. (For economy of space, the figure only shows vision.) Each modality delivers a limited amount of information at a particular rate to Working Memory. On the bottom left-hand side of the figure is the Motor Processor or Motor, while on the right-hand side of the figure is Cognition, which consists of Working Memory, Long-Term Memory, and the Cognitive Processor. In the MHP, tasks are decomposable into deliberate actions that take place in the form of discrete sequential operator applications in Cognition. Some operators may require input from the Perception, others may initiate external acts using Motor, and finally, some may use existing data in Working Memory for internal calculations that are independent of Perception and Motor.

Each subsystem in the MHP has a range of nominal cycle times. The cycle times that we use in our model for Perception and Motor are MHP median figures. The cycle time we use for Cognition is taken from John (1988) and is 20 msec faster than the MHP median figure for Cognition, but well within the range of nominal cycle times for Cognition in the MHP.

Perceptual (vision) cycle time (τ_p)	= 100 msec	(Range: 50-200 msec)
Cognitive cycle time (τ_c)	= 50 msec	(Range: 25-170 msec)
Motor cycle time (τ_m)	= 70 msec	(Range: 30-100 msec)

The total time required to perform a task is calculated by first creating an algorithm for performing the task using operators. Using the cycle times of each processor to compute the duration of each operator, the total time to complete the algorithm can be calculated. The three processors can run in parallel, so, for instance, once a motor command has been initiated, other cognitive operators can be applied at the same time, as long as they are not dependent on the result of the motor operator.

*This research was sponsored by grant NCC2-517 from NASA Ames.

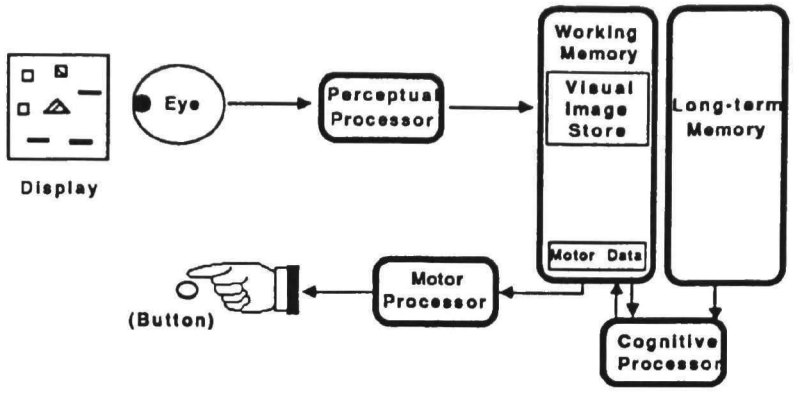

Figure 1: The Model Human Processor

We are using Soar [Laird *et al.*, 1987] as the underlying architecture for implementing our operator sequences. Soar has all of the basic subsystems of the MHP, including Perception, Cognition, and Motor.[1] In Soar there is a Working Memory and a Long-term Memory, and deliberately selected operators provide the basis of action. Soar's Long-term Memory is a parallel production system. Perception is implemented in Soar as Lisp functions that transduce environmental stimuli and send input to working memory, while Motor is implemented in Soar as Lisp functions that receive output from Working Memory and then act on the environment. In Soar, low-level Perception occurs in parallel with the firing of productions, as does Motor, once it has been initiated through an operator application.

The basic deliberative act in Soar is the selection of an operator to apply. Thus the operators in the MHP map directly onto operators in Soar. Both the selection and application of operators are performed by productions matching against Working Memory and suggesting changes to it. Thus, productions control which operators are selected, and once selected, how they apply. In Soar, specific operators (that is, instantiations of operator types) are created as soon as the data needed to instantiate them are available. In general, this means that new operators will be created, and ready for selection, during the application of the currently selected operator. However, only one operator can be applied on any given cycle. Although Soar is programmed at the level of productions, operators are the appropriate level for describing algorithms that correspond to processing in the MHP.

2 A Model of Visual Attention

The goal of our work is to develop a complete, computational model of visual attention. Unfortunately, there is no existing computational theory which covers all visual attention phenomena. There are many experimental paradigms with accompanying data that provide constraints for a complete model. In this paper we will model three general phenomena: humans are able to control the portion of the visual field that they attend to [Sperling, 1960]; attention is required for object identification [Treisman and Gelade, 1980; Treisman and Schmidt, 1982]; and precuing facilitates object identification [Colegate *et al.*, 1973]. These are critical visual attention phenomena which are all observed in letter-wheel task of CHE. We will use this task to demonstrate the details of the model. Our implementation in Soar has also been used to simulate other phenomena that have been attributed to visual attention, including illusory conjunctions and various reaction time behaviors associated with search [Wiesmeyer and Laird, 1990], that are beyond the scope of this paper.

Figure 2 shows how we have extended the MHP to include visual attention. On the top left-hand side of the figure again is Perception, but it has now been expanded to show details of how it delivers visual data to Working Memory. A single ovoid *region of attention* controlled by Cognition, as in Eriksen's "Zoom Lens

[1] Soar has additional capabilities for planning and learning that are not needed for the tasks described in this paper.

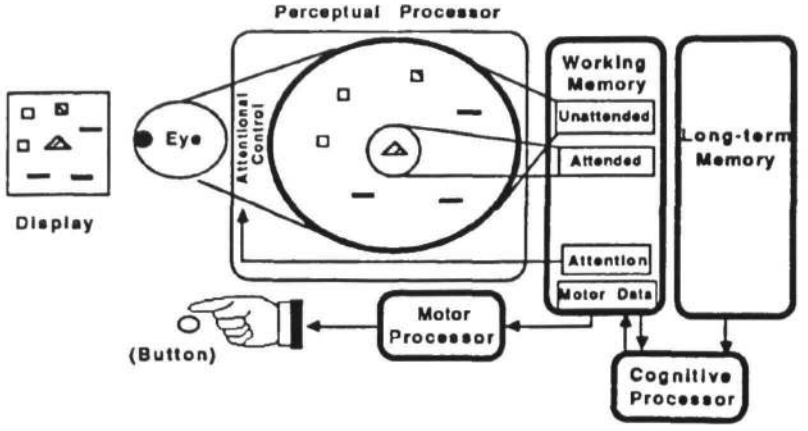

Figure 2: The model human processor with visual attention extension

Theory" [Eriksen and Yeh, 1985; Eriksen and St. James, 1986], separates the visual field in Perception into two regions: attended and unattended. The region of attention is under control of Cognition and can be moved around the visual field. Perception, as in the MHP, converts stimuli from these regions into symbols that are delivered to Working Memory; we call these symbols *features* after Treisman [Treisman and Gelade, 1980; Treisman, 1987]. Thus, input to the Visual Image Store in Working Memory is separated into two sets: *attended features* from the attended region and *unattended features* from the unattended region.

In our model, features are of two types: shape and color, although color is not needed to model the letter wheel task. Motion (change) information is associated with features, when appropriate, and features can be *marked* by productions. Marking allows for a simple memory of whether a feature has been attended. Except for motion information and marks, features have no other explicit properties besides their identities.

Attention is controlled through the `Feature-Shift` operator. A separate operator is created for each unmarked feature in the visual field and application shifts the region of attention to that feature. In shifting attention, the new region of attention may also include other features that are near the selected feature. The `Feature-Shift` operator can be further specialized to shift to the *nearest* feature of a given type. For example, `Feature-Shift` can be used to shift to the nearest new shape feature.

An important property of our model is that attention improves the resolution of shape features. We have adopted a simplified version of Eriksen's "Zoom Lens Model" of attention, which theorizes that unattended stimuli are always lower in resolution than attended stimuli and that there is an inverse relationship of region extent and attended feature resolution. Low resolution means, for instance, that a letter stimulus that does not appear in the region of attention or appears in a relatively large region of attention, might appear as a unidentifiable "blob" in Working Memory.

3 The Letter Wheel Experiment

We are using Colegate, Hoffman and Eriksen (1973) as a prototypical example of an object identification task that involves precuing. Other experiments have similar results under slightly different conditions [Eriksen and Hoffman, 1972; Eriksen and Hoffman, 1973]. In CHE, subjects were presented with displays similar to those shown in Figure 3. The task was to identify the cued letter. Subjects responded vocally. Letters from the set `A,E,M,U` were systematically distributed over the various locations in 8 and 12 letter displays so that the effects of letters neighboring the target on reaction time could be studied. Displays were 2 degrees in visual angle and presented with a tachistoscope—2 degrees is the extent of the fovea and approximately the diameter of the entire letter wheel on the right of the figure if this paper is held at arm's length.

There were two basic modes of stimulus delivery, simultaneous presentation and precued presentation. In both modes subjects were told to fixate on a fixation point until a cue for the location of a letter to be identified was presented. In simultaneous presentation the cue and the letter arrived at the same time, while

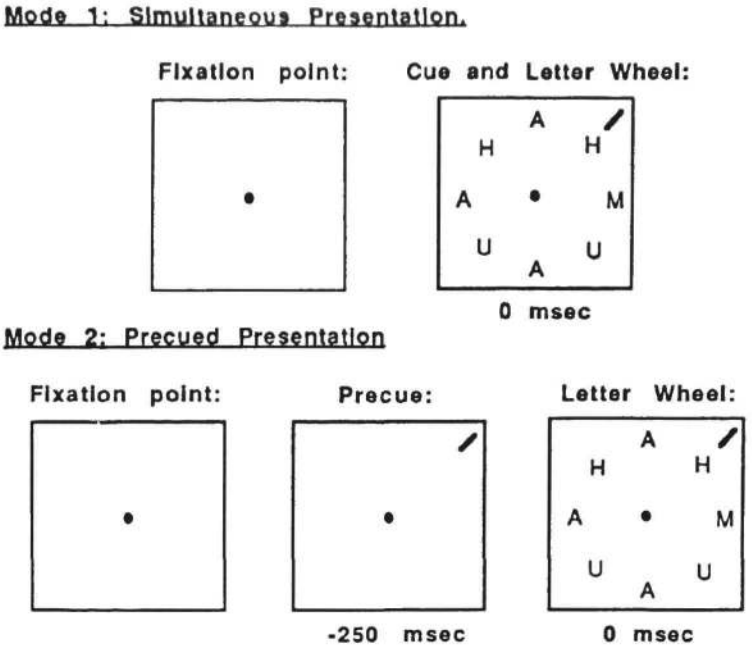

Figure 3: Example letter wheel stimuli

in precued presentation the cue preceded the letter wheel by an interval of up to 350 msec. In both modes, subjects were required to identify the cued letter. CHE's data showed that reaction times improve at a constant rate as the precue period increases up to 250 msec. With a 250 msec precue there was a 100 msec improvement in reaction time with respect to the simultaneous presentation condition. Response times for 8-letter displays were always slightly faster than for 12-letter displays. We do not account for the differences between the 8 and 12 letter displays in this paper, but have concentrated on getting the response times of our simulations within the range of observed data.

4 Modeling the Letter Wheel Experiment

To model these experiments requires three operators in addition to `Feature-Shift`:

`Identify` : Identifies an attended object based on the shape feature (cue and target in our simulations).
`Verify` : Verifies that an object identified (the cue in our simulations) was the one expected.
`Respond` : Performs the appropriate motor action.

These operators form the minimal set we could find to perform the task given the level of computation assumed to be possible in the supporting productions. The only questionable one might be `Verify`. This is used to ensure that the result of a `Feature-Shift` and subsequent `Identify` was correct. Without it, the system would be more prone to errors.

We assume that τ_c is the time required for an average cognitive operator to be created or selected and applied—creation being separate from selection and application; τ_p is the time required for a stimulus to get from the retina to working memory; and τ_m is the time required for a new Working Memory element to effect a motor command. Thus `Identify` and `Verify` take 100 msec for creation, selection, and application, while `Respond` takes an additional 70 msec (170 msec total) because of the motor cycle time.

The application of `Feature-Shift` involves interaction between Cognition and Perception, which makes determining its time course problematic. We decompose its time course into the three stages required for Cognition to (1) react to a new stimulus, (2) shift visual attention to that new stimulus, and (3) then receive new features in Working Memory that reflect the change in visual attention. The only unknown component of this series of events is the time required for a visual stimulus to travel from the locus of action of attention

to Working Memory; it can be estimated from brain data: Visual attention has been shown to act at a location in the brain called V4 [Moran and Desimone, 1985]. From V4 attended stimuli go to the inferior temporal cortex and then to the amygdala and the hippocampus [Mishkin and Appenzeller, 1987]. The amygdala and hippocampus are known to be involved in visual memory and are thus assumed to be the beginnings of Cognition (Working Memory). Since V4 is physically located about half way between the retina and the amygdala and hippocampus, we estimate that it should take about 1/2 of τ_p (or 50 msec) for the effects of visual attention to be reflected in the contents of Working Memory. Thus, we estimate the amount of time required for the entire process as:

100 msec	τ_p for propagation of data from the retina to working memory
50 msec	τ_c for `Feature-Shift` creation
50 msec	τ_c for `Feature-Shift` selection and application
50 msec	$\frac{1}{2}\tau_p$ for propagation of attended data from V4 to working memory
250 msec total	

For features already in Working Memory, the time required to shift attention to them and then receive new data reflecting the change in attention is the sum of the last three steps above or 150 msec. Admittedly, it would be preferable to use an empirically derived time in our model rather than these estimates, but we have not been able to locate a suitable one in the literature and have found this time to work well in our algorithms.

Bearing these figures in mind and the fact that next operator creation can overlap with current operator application, the job at hand in creating an algorithmic cognitive model is to fit operators together so that the requisite task may be done and known time constraints may be satisfied.

We have developed algorithms for both the simultaneous presentation and 250 msec precue conditions. Observed mean response times from CHE are 574 msec (483 - 635 msec) for the simultaneous presentation condition and 491 msec (427 - 569 msec) for the 250 msec precued presentation condition. Our algorithms predict 570 msec and 470 msec, respectively, and are thus both within the range of times obtained from experimental subjects. Additionally, the improvements due to precuing predicted by our model (100 msec) correlate closely with those of actual subjects (mean 83 msec)[2]

Details of the both algorithms follow. Both are implemented in Soar using a single set of productions that controls the selection of operators based on the contents of Working Memory—there is no explicit selection strategy programmed for the two tasks. In these traces, the selection of operators is implied and occurs for each operator immediately preceding its application.

In the simultaneous presentation trace in Figure 4, a single `Feature-Shift` is performed to the cue. Both the cue and the letter are in the region of attention and their features are in the new attentional region at 200 msec. The algorithm next identifies and verifies that the cue is in the attentional region (250 - 400 msec) and then identifies the letter (400 - 450 msec) and then responds (450 - 570 msec).

In the 250 msec precue case shown in Figure 5, a `Feature-Shift` is applied to shift to the cue. From start to finish, this takes 250 msec (-250 - 0 msec). Following the shift, the cue is identified and verified (0 - 100 msec). At the same time, the letter wheel is being processed by perception and becomes available in Working Memory at 100 msec. At this point, a second shift is performed to the nearest letter (150 - 250 msec) followed by its identification (250 - 300 msec) and finally the response (350 - 470 msec).

In the simultaneous presentation condition, only one shift of attention is required, while in the precued condition two shifts are required. In spite of this, the precued condition is still faster. There are two reasons for this: (1) The first shift of attention for the precued condition finishes before the wheel is presented and timing starts. (2) The `Identify` and `Verify` operators for the cue occur during the period in which the wheel stimulus is traveling from the retina to working memory (labeled "overlap" in the second algorithm).

The CHE results show that there is no improvement in reaction time when precuing exceeds 250 msec. The response time of our implementation does not improve either, since the shift to the target letter is contingent upon its shape feature appearing in Working Memory, which time will never change from 100 msec after the stimulus is presented.

[2]There is a great deal of variation in experimental results for precuing in object identification experiments. The CHE experiment was chosen simply on the basis of its typicality.

Time	Operator Event	Event/Comment
		Attention centered on fixation point
0 msec	None	Wheel and cue at retina
50 msec	None	Wheel and cue in P
100 msec	Feature-Shift created	Wheel and cue in WM
150 msec	Feature-Shift applies	Shift to cue and letter
200 msec	Feature-Shift completed	Attentional message at V4
250 msec	Identify (cue) created Identify (letter) created	Cue and letter attended in WM
300 msec	Identify (cue) applies Verify (cue) created	
350 msec	Verify (cue) applies	
400 msec	Identify (letter) applies Respond created	Letter identified
450 msec	Respond applies	
500 msec	Respond completed	Motor output begins
570 msec		Motor command completed

Figure 4: Operator trace of the simultaneous presentation of the letter wheel and cue

The results of the model for precuing between 0 and 250 msec are less clear. The CHE results show there is incremental improvement as the precue goes from 0 to 250 msec (in 50 msec increments). Between 0 and 150 msec the predictions of our model are unclear because the letter wheel is being processed by Perception while attention is being shifted. If we assume that the time it takes Perception to process data is stochastic, then the increase in precue would increase the probability that the precue would be available soon enough to allow the extra shift.

5 Discussion

In this paper, we have developed a computational model of visual attention and algorithms that are sufficient for calculating the reaction times of a simple task that requires visual attention. In the process, we have estimated the time for Cognition to react to a new stimulus, shift visual attention to that new stimulus, and receive new features in Working Memory that reflect the change in visual attention.

The exact results of this model are dependent on many assumptions including the cycle times for Perception, Cognition, Motor, and Attention. However, these assumptions are not unique to this model; all except the time for changes in attention were based on previous research. The set of operators that we have chosen is another key assumption. A challenge for our future modeling is to see if these operators are sufficient for additional tasks requiring visual attention.

Time	Operator Event	Event/Comment
		Attention centered on fixation point
-250 msec	None	Cue at retina
-200 msec	None	Cue in P
-150 msec	Feature-Shift created	Cue in WM
-100 msec	Feature-Shift applies	Shift to cue
-50 msec	Feature-Shift completed	Attentional message at V4
0 msec	Identify (cue) created	Cue attended in WM / Wheel at retina
50 msec	Identify (cue) applies Verify (cue) created	Wheel in P
100 msec	Verify (cue) applies Feature-Shift(s) created	Wheel in WM
150 msec	Feature-Shift applies	Shift to nearest letter
200 msec	Feature-Shift completed	Attentional message at V4
250 msec	Identify (letter) created	Letter attended in WM
300 msec	Identify (letter) applies Respond created	Letter identified
350 msec	Respond applies	
400 msec	Respond completed	Motor output begins
470 msec		Motor command completed

Figure 5: Operator trace of the 250 msec precue

A major simplification in the current implementation is that it is unnecessary to model the inverse relationship between region size and shape stimulus resolution strictly. Instead, we assumed that unattended shape features were simply not identifiable, but they were detectable. This allowed them to act as destinations in the visual field for shifts of attention. Once a feature was attended, its identity was available through Working Memory; that is, an `Identify` operator could be created for it. We may need to refine our interpretation of the "Zoom Lens Model" so that identifiability takes on a more stochastic quality based on proximity to the region of attention instead of the discrete quality it now has. It is possible that this stochastic quality could help to account for noise effects, such as those found in CHE, where distracting letters near the target slowed identification response time.

Although extensions are necessary, the results of this work show that the augmented MHP can be used to account for a limited group of visual attentional phenomena, and that Soar is a sufficient symbolic architecture for building running MHP models.

References

[Card *et al.*, 1983] S.K. Card, T.P. Moran, and A. Newell. *The Psychology of Human-Computer Interaction.* Erlbaum, Hillsdale, NJ, 1983.

[Colegate *et al.*, 1973] R. Colegate, J.E. Hoffman, and C.W. Eriksen. Selective encoding from multielement visual displays. *Perception and Psychophysics*, 14:217–224, 1973.

[Eriksen and Hoffman, 1972] C.W. Eriksen and J.E. Hoffman. Some characteristics of selective attention in visual perception determined by vocal reaction time. *Perception and Psychophysics*, 11:169–171, 1972.

[Eriksen and Hoffman, 1973] C.W. Eriksen and J.E. Hoffman. The extent of processing noise elements during selective coding from visual displays. *Perception and Psychophysics*, 14:155–160, 1973.

[Eriksen and St. James, 1986] C.W. Eriksen and J.D. St. James. Visual attention within and around the field of focal attention: A zoom lens model. *Perception and Psychophysics*, 40:225–240, 1986.

[Eriksen and Yeh, 1985] C.W. Eriksen and Y.Y. Yeh. Allocation of attention in the visual field. *Journal of Experimental Psychology*, 11:583–597, 1985.

[John and Newell, 1989] B.E. John and A. Newell. Cumulating the science of HCI: From S-R compatibility to transcription typing. In *Proceedings of CHI*, pages 109–114, Austin, Texas, April 30-May 4 1989.

[John, 1988] B.E. John. *Contributions to engineering models of human-computer interaction.* PhD thesis, Carnegie Mellon University, 1988.

[Laird *et al.*, 1987] J. E. Laird, A. Newell, and P. S. Rosenbloom. Soar: An architecture for general intelligence. *Artificial Intelligence*, 33(3), 1987.

[Laird *et al.*, 1990] J.E. Laird, K. Swedlow, E. Altman, and C.B. Congdon. Soar 5 user's manual. Technical report, University of Michigan, 1990. In preparation.

[Mishkin and Appenzeller, 1987] M. Mishkin and T. Appenzeller. The anatomy of memory. *Scientific American*, June:80–89, 1987.

[Moran and Desimone, 1985] J. Moran and R. Desimone. Selective attention gates visual processing in the extrastriate cortex. *Science*, 229:782–784, 1985.

[Newell, 1990] A. Newell. *Unified Theories of Cognition.* Harvard University Press, Cambridge, MA, 1990. (in press).

[Sperling, 1960] G. Sperling. The information available in brief visual presentations. *Psychological Monographs*, 74, 1960.

[Treisman and Gelade, 1980] A. Treisman and G. Gelade. A feature integration theory of attention. *Cognitive Psychology*, 12:97–136, 1980.

[Treisman and Schmidt, 1982] A. Treisman and H. Schmidt. Illusory conjunctions in the perception of objects. *Cognitive Psychology*, 14:107–141, 1982.

[Treisman, 1987] A. Treisman. Properties, parts, and objects. In K. R. Boff, L. Kaufman, and J. P. Thomas, editors, *The Handbook of perception and human performance.* Wiley-Interscience, New York, 1987.

[Wiesmeyer and Laird, 1990] M.D. Wiesmeyer and J.E. Laird. A computer model of visual search. Technical report, Artificial Intelligence Laboratory, The University of Michigan, March 1990.