ELSEVIER

Artificial Intelligence 75 (1995) 135–160

Artificial Intelligence

Behaviorist intelligence and the scaling problem

John K. Tsotsos[*,1]

Department of Computer Science, 6 King's College Rd., University of Toronto, Toronto, Ontario M5S 1A4, Canada

Received October 1992; revised November 1993

Abstract

This paper argues that the *strict* computational behaviorist position for the modeling of intelligence does not scale to human-like problems and performance. This is accomplished by showing that the task of visual search can be viewed within the behaviorist framework and that the ability to search images (or any other sensory field) of the world to find stimuli on which to act is a necessary component of any behaving, intelligent agent. If targets are not explicitly known and used to help optimize search, the search problem is NP-hard. Knowledge of the target is of course explicitly forbidden in the strict interpretation of the published behaviorist dogma. Also, the paper summarizes the existing neurobiological and behavioral realities as they pertain to behaviorist claims. The conclusion is that there is very little support from biology for strict behaviorism. Strict adherence to the philosophy of the behaviorists means that efforts to demonstrate that the paradigm scales to human-size problems are certain to fail, as are attempts to evaluate it as a model of human intelligence. The strict position thus cannot be what the behaviorists really mean. It would benefit the research community if they could elucidate their terms, and provide theoretical arguments that support claims of scalability.

1. Introduction

Recently, the philosophy for realizing intelligent behaviors in machines as articulated by Brooks, his colleagues and others has received a great deal of attention and has attracted many vocal supporters as well as equally vocal

* Telephone: 416-978-3619. Fax: 416-978-1455. E-mail: tsotsos@vis.toronto.edu.
[1] CP-United Fellow of the Canadian Institute for Advanced Research.

SSDI 0004-3702(94)00019-W

detractors (Brooks [6–8], Aloimonos and Rosenfeld [1], Ballard [3]).[2] Brooks claims

> . . . that intelligence be reactive to dynamic aspects of the environment, that a mobile robot operate on time scales similar to those of animals and humans, and that intelligence be able to generate robust behavior in the face of uncertain sensors, an unpredicted environment and a changing world. . . (R. Brooks, Computers and Thought Lecture, International Joint Conference on Artificial Intelligence, Sydney, August 1991)

It seems difficult and unreasonable to argue against this philosophy, yet many do. Another of Brooks' beliefs is that machines constructed out of simple modules with simple communication will exhibit intelligent behavior as an emergent property; the behavior is not directed by a single homunculus nor is it explicitly specified in the machine in any way. This too sounds like it is an approach to complex behavior that is worth pursuing.

1.1. Setting up the controversy

Where is the controversy? The cornerstones of the "subsumption" architecture Brooks proposed in [5] for intelligent control include: control layers define a total order on a robot's behaviors; the dominance of layers follow a hypothesized evolutionary sequence; each layer may "spy" on layers at lower levels and "inject" signals into them. It is claimed that the structure is scalable to human-like behavior[3] and Brooks argues strongly against many of the currently prominent concepts and activities in AI. Specifically, he believes that there is no place for: the sense-model-plan-act framework for robot control; the representation of intermediate[4] results, the use of hierarchical computations; the explicit representation of goals; and, CAD-like models of the world. As proof of his position he offers evidence that is compelling: many mobile robots that seem to have robust and interesting performance (Allen, Herbert, Tom & Jerry, Attila, Squirt, Allenmore, and others).

Brooks seems to be re-kindling the torch of old behaviorism, a philosophy appearing about 1913 in the psychology community (see Watson [49], the "founder" of the behaviorist position in psychology). Behaviorism stood for one basic belief: humans are biological machines and as such do not consciously act, do not have their actions determined by thoughts, feelings, intentions or mental

---

[2] In addition, proponents have received two prestigious awards: Rod Brooks was a co-recipient of the 1991 IJCAI Computers and Thought Award and Dana Ballard was awarded the 1989 IJCAI Best Paper Prize.

[3] "I think that the new approach can be extended to cover the whole story, both with regards to building intelligent systems and to understanding human intelligence". (Brooks [7, p. 585].

[4] For this paper, an intermediate representation is one which has the following qualities: it can be considered as an input representation by two or more other processes; it is not a representation of the input in its raw form.

processes. Human behavior is a product of conditioning: humans react to stimuli. Stimulus-driven internal states were allowed, but internal mental states were not. Arguments against behaviorism are easily found in the cognitive science literature.[5]

Similarly, arguments against Brooks' position are not new. Kirsh, for example, focuses on one of Brooks' claims, that intelligent behavior is "concept-free" [26]. Kirsh claims that concepts are necessary for some types of behavior and also can make computational processes simpler. He argues for the need of representation in a theory of perception simply because vision is complex and must be sometimes solved in general ways.

Three other papers present sympathetic and complementary views to that of Brooks. Aloimonos and Rosenfeld coined the term "purposive and qualitative vision" and claim that the "purpose" of a vision system was neglected in past research [1]. They may have forgotten to note that the vision systems that many tested in the mid-to-late 1970s and early 1980s, although they may have had other failings, did not have the failing of ignoring purpose and qualitative descriptions (see [47] for a comprehensive review). On the contrary, those systems were very much task-oriented and model-based; their outputs were typically qualitative, task-specific representations of the scenes using some form of natural language-like primitives. In fact, the focus on task-specific solutions was the major criticism leveled against that work. They further state that for many vision problems, complete and accurate recovery of the scene is not necessary. This is an amazing understatement. In fact, not only is it not necessary, it is not a computationally tractable problem [43, 44]. Aloimonos and Rosenfeld basically support Brooks' stand because they propose that research concentrate on specific vision-guided behaviors.

Ballard also proposes a similar view but more focused on visual behavior; he uses the term "animate vision" [3]. He claims that animate vision systems must have gaze control and thus can be vastly less expensive when considered in the large context of behavior. This is far from obvious; in fact, there are many cases where it is not true [48]. Agreeing with Brooks, Ballard says that elaborate categorical representations may not be needed. Further, he claims that memory is not required since an animate vision system can compute a representation of the vision world rapidly on demand. This last claim (if not also the previous ones) is easily refuted.

The above positions will be grouped together for the purposes of this paper since they contribute to "computational behaviorism" (for the remainder of the paper, the term behaviorism is intended to refer to the new computational version as opposed to the old psychology position). One further theory will be briefly mentioned since it is often used as evidence from the biological community for the

[5] For example, the following appeared in a recent overview of cognitive neuroscience. Kandel and Squire point out that "It was easy to show that any study of mental activity that failed to consider representations of mental events was inadequate to account for all but the simplest forms of behavior" [24, p. 143].

above computational paradigms; it should be clear that the following work does not fall into the "psychological behaviorist" camp. Nevertheless, Ramachandran's [36] utilitarian theory is remarkably similar to the positions outlined so far. Ramachandran rejects previous well-known theories (Helmholtz's perception as unconscious inference, Gibson's direct perception, Marr's natural computation) and proposes rather that perception does not involve intelligent reasoning, nor resonance with the world, nor the creation of internal representations. Rather, perception is a *bag of tricks*. Through millions of years of evolution, the visual system has evolved numerous short-cuts, rules-of-thumb, and heuristics each one adopted only because it works and not because of any other appeal. However, some stimulus must be responsible for triggering or activating the various tricks at the appropriate times. The trigger is not necessarily the stimulus itself, but may be some early representation of the stimulus that is extracted in a mechanical fashion. But all of the tricks cannot be always active; there would be too many and it would imply that spatial parallelism is sufficient for perception (arguments against this position will be recounted later). The processing that is required for stimulus recognition is of the same kind as required for stimulus recognition in the behaviorist schemes above. The key problem with Ramachandran's utilitarian theory is not in the early processing but rather that it is insufficient with respect to the control of which tricks can be active and which are executed at a given point of time.

This paper will focus on two claims of the computational behaviorist's philosophy which are difficult to prove: that behaviorism will scale up to problems which are human-like in their size; and, that behaviorism can be used as a model of human intelligence. It is important to present a definition of "scale up to human-like problems" in order to make the discussion concrete. A computational theory scales to human-like problem sizes if:

- the algorithm that embodies the theory accepts up to the same number of input samples of the world per unit time as human sensory organs,[6]
- the implementation that realizes the algorithm exists in the real world and requires amounts of physical resources which exist,
- the output behavior of the implementation as a result of those stimuli is comparable both in quality and timing to human behavior[7] (i.e., it would be indistinguishable from human behavior in all important respects).

For the remainder of the paper, the term "scales up" will be grounded in this definition. Many seek escape from the difficulties inherent in dealing with human behavioral performance by claiming their models are not intended to be models of

---

[6] It is a nontrivial task to determine exactly the quantitative nature of the input to the human sensory system. With respect to the visual system, there are two eyes; each has about 110–125 million rods and 6.3–6.8 million cones; each eye can discriminate over a luminance span of 10 billion to one; the spatial resolution of the system peaks at about 40 cycles/degree while the temporal resolution peaks at about 40 Hz but the two are not independent; finally, there are many inputs from other sensory and motor areas. See [13, 14] for further discussion. Similar data is needed for non-visual sensory systems.

[7] The behavioral literature on exactly what the quality, quantity and timing of human behavior is to a variety of stimuli is immense, but far from complete.

human intelligence (Ramachandran, Ballard and Brooks, however, do claim biological plausibility). Note that the definition does not necessarily require that either the algorithm or its implementation have any relationship to how humans process information, nor is there any biological restriction on the amount of resources used (numbers of processors, for example, are not restricted to be less than the number of neurons in the brain). Thus, there is no implication that this definition applies only to models of biological behavior; it is just a specification of problem size and system performance. However, for those researchers who claim biological plausibility, the following must be added to the definition in order to define "scaling with biological plausibility":

- solutions should require significantly fewer than about 10^{12} processors operating in parallel if modeling the whole brain (the visual cortex of the macaque monkey is about 60% of the cortex; the figure above includes the cortex as well as the various sub-cortical structures [16]), each able to perform one weighted sum computation over all its input per millisecond (possibly with an added nonlinearity such as thresholding);
- processor average fan-in and fan-out should be about 1000 overall[8] so that the total number of connections is on the order of 10^{15} [24]; and
- solutions should not involve more than a few hundred sequential processing steps (the firsts spikes arrive in inferotemporal cortex about 80–100 ms after the onset of a stimulus in the monkey with resulting task-driven eye movements initiated about 250 ms after the stimulus onset [9].

The arguments behaviorists present on scaling are inadequate (for example, see [6, 7]). Recently, a special issue of *SIGART Bulletin* [37] included the proceedings from the Workshop on Integrated Cognitive Architectures, and a large number of presentations were either motivated by, or support, the behaviorist position. The proposed schemes fail on issues of scalability and cognitive plausibility. This issues contains at least 15 papers on work that attempts to integrate perception with action. All of the authors claim that their methods scale nicely. Almost all of them trivialize the perception component, assume they are given correct and abstracted input or say nothing about how perception is to be solved. All consider the issue of cognitive plausibility in a very superficial manner. One of the papers (Ogasawara, p. 140) at least attempts a complexity analysis to support claims for scalability. Unfortunately, if the issues of scaling to human-like problems plus cognitive plausibility are taken seriously, even polynomial algorithms are too slow (see [44–46]).

1.2. A different perspective on the behaviorist strategy

One can view the behaviorist intelligence paradigm as a particular implementation of one of the most basic tools of AI: the well-known "hypothesize-and-test"

[8] Connectivity varies with cell type: stellate cells may receive a few hundred synapses, small pyramidal cells a few thousand, while large pyramidal cells may receive tens of thousands of synapses. Most if not all computational models of perception use the average figure of 1000 as a guide.

strategy for search. It is easy to place the hypothesize-and-test idea into a reactive framework: assume (as is done in such frameworks) that the set of choices of stimulus–action pairings is given. The behavior specification of the device can easily provide this. The standard hypothesize-and-test paradigm operates as follows:

- Acquire the current input and determine which aspects of it are salient to the problem.
- Propose a particular explanation (hypothesis) as the correct one for the current input.
- Devise a test in order to verify that it is indeed the correct explanation.
- If the hypothesis passes the test, proceed with that explanation and its consequences. If the hypothesis fails the test, select another explanation and try again. The selection mechanism may be quite complex depending on the size of the hypothesis space.

Suppose now that all hypotheses can be tested in parallel. Connell [10] for example, defines nodes that have the above functionality. This would lead to a configuration such as shown in Fig. 1; this is not unlike the kinds of circuits derived using subsumption ideas. Fig. 1 is an abstraction of the several control diagrams given in [10]. The stimulus transducer passes to each stimulus–action pair the subset of input data that is relevant for that action. The node stimulus–action determines whether or not the stimulus can indeed act as a trigger for the behavior it represents. Since at this stage, the triggered behavior has not yet received final routing to the robot's actuators, it may be considered as one of several hypotheses that must be considered as active candidates for response to

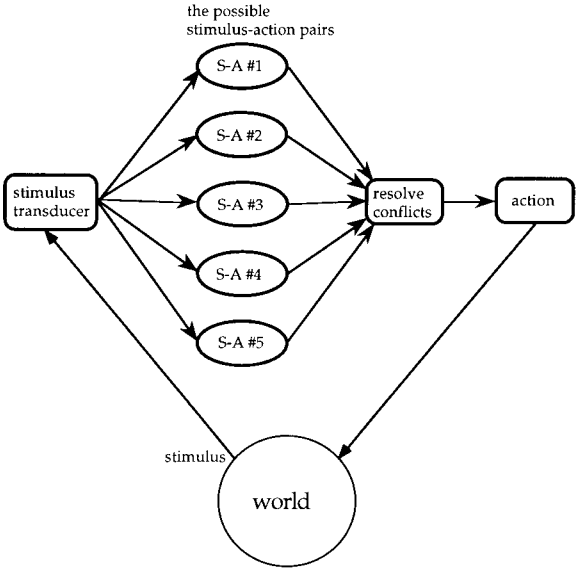

Fig. 1. An abstraction of Connell's architecture.

the current stimulus. Hypotheses are thus represented and tested individually and in parallel. Actions are executed that move the sensors or do whatever else is appropriate depending on the best hypotheses after a conflict resolution step. This circuit is reactive: it reacts to stimulus as they enter the system, and it appears to have many behaviors since it has many stimulus–action pairs to choose from and it resolves conflicts. But most importantly, it has exactly the same hypothesize-and-test mechanism that has been common in AI.[9]

The key difference is the parallel implementation. This should have triggered a warning however: recall Neisser's claim that a theory of perception based on spatial parallelism alone is quantitatively inadequate to explain human perceptual behavior [33]. Effective quantitative demonstration of this claim was provided in [42]. This extends easily to behavior in general since behavior is stimulus-driven. In effect, this is exactly what Brooks is proposing: that parallel hypothesize-and-test of stimulus–action pairs is sufficient to cause an agent to exhibit intelligent behavior. This is feasible only for relatively small stimulus–action pair spaces, such as the ones which are currently implemented in various reactive devices. Interestingly, one of the behaviorists, Mataric, argues that the behaviorist's strategy is not appropriate for high bandwidth, high density sensors such as vision [27]. She points out that coarse grain, low density, low bandwidth sensors naturally lead to solutions involving parallelism for stimulus–action activation.

2. Perception and the scaling problem

The scaling problem has been addressed by many researchers in the context of specific algorithms for specific visual problems. The analysis below follows [43, 44, 47, 48]. The approach here is to look at vision "in the large" by considering the problem of visual search, a ubiquitous activity within intelligent behavior. Visual search can be viewed within the behaviorist paradigm and is an integral part of any intelligent behavior. In a typical experimental setup for the study of visual search behavior in humans, a subject is presented with a target (or targets), and a test image and asked to determine whether or not that target is present in a test image. This involves two behaviors (or stimulus–action pairs): (1) if target present, press button A; (2) if target absent, press button B. Thus, the task has exactly the same characteristics as shown in Fig. 1, with two possible stimulus–action pairs.

It is difficult to imagine any vision system which does not involve similar visual search operations, and it is clear that these types of operations appear from the earliest levels of vision systems to the highest. Vision usually begins with a set of measurements at a number of locations in a sensory field (pixels). In most cases, the first step of visual information processing is to create an edge representation. What are the physical structures responsible for the edges? Edge-finding is an

[9] This view of the behaviorist strategy first appeared in a discussion relating the behaviorist paradigm to the concept of incremental active vision in [47, 48].

instance of visual search: given a model of an edge, is there an instance of this edge model in the test image? This search problem is *unbounded*: there are no a priori constraints on the size, extent, shape, etc. of the structures sought. The model base of possible objects in the world would be of no help since it would be very large and varied in general and thus not provide useful overall constraints (all possible edge types could exist at all possible image locations). Region growing, shape matching, structure from motion, the general alignment problem, and connectionist recognition procedures, etc., are also specialized versions of visual search in that the algorithms must determine which subset of pixels corresponds to a given prototype or description or satisfies some set of constraints. This search task is precisely what any model-based computer vision system also has as its goal: given a target or set of targets (models), is there an instance of a target in the test display? Visual search tasks are integral components of any visual information process.

2.1. Visual search

According to the definition provided by Rabbitt [35] visual search is a categorization task in which a subject must distinguish between at least two classes of signals: goal signals which must be located and reported and background signals which must be ignored. In [43], a computational definition of the visual search task was presented that contained two important subcases: unbounded visual search in which either the target is explicitly unknown in advance or it is somehow not used in the execution of the search; and bounded visual search, in which the target is explicitly known in advance in some form that enables explicit bounds to be determined that can be used to limit the search process. These bounds may be in the form of spatial extent of the target, feature dimensions that are involved or specific feature values. Then, a proof was given showing that the unbounded case is NP-complete in the size of the image, while the bounded case has linear time complexity in the same variable. The proofs are not specific to vision; they apply equally well to search problems with any modality of sensory information. The NP-completeness of the unbounded case is due solely to the inability to predict which pixels of a test image correspond to objects in a non-exponential manner. From the behaviorists' perspective, the ability to find stimuli on which to act is necessary, and thus the results of those theorems apply directly to behaviorism.

Let a test image I contain p pixels, and there are M feature values associated with each pixel. Thus, $|I| = pM$. The size of the target image T is defined in a similar manner, $|T| = qn$; the values of q and n corresponding to the number of image locations and number of feature values represented in the target respectively. An instance of the Visual Search problem is specified as follows:

- a test image I,
- a target image T,
- a difference function $diff(a)$ for $a \in I$, $diff(a) \in \mathbb{R}_\rho^0$, ($\mathbb{R}_\rho^0$ is the set of non-negative real numbers of fixed precision ρ),

- a correlation function $corr(a)$ for $a \in I$, $corr(a) \in \mathbb{R}_{\rho}^{0}$,
- two thresholds, θ and ϕ, both positive integers.

The *diff* and *corr* functions basically constrain the acceptability of a solution by requiring that the solution subset have sufficiently low error and be the maximal subset satisfying the error measure respectively. The functions may be any appropriate functions (even table lookup). They are only constrained to require at most polynomial time complexity to compute. For example, the *diff* function may compute an L_1 error norm, while the *corr* function may compute a correlation. Images do not refer to retinotopic collections of pixel elements only. Rather, images can be thought of as abstractions ("intrinsic images") or as collections of features indexed by image location. More detail may be found in [43]. It must be emphasized that this formalism is not necessarily intended as an implementation level description of the problem (although it could be implemented exactly as stated). It is an "in principle" solution, to use Marr's levels of abstraction, at the computational level. Both passive (fixed stimulus acquisition system) and active or animate (dynamic, feedback-controlled stimulus acquisition system) situations are addressed by the work.

Four problems will now be given with basic theorems regarding their time complexity. The proofs and detailed discussions are found in [43, 47, 48] as are the intuitions behind their definitions and relationships to real-world vision problems. These discussions are not repeated here; however, the theorems are re-stated both for completeness and because they form the basis of the new theorems presented in Section 3 and the discussion of Section 4.

2.1.1. Unbounded passive visual search (UPVS)

In the unbounded version of visual search, no guidance is permitted and search proceeds blindly as in the following formulation. Given a test image, a difference function, and a correlation function, is there a subset of pixels of the test image such that the difference value is less than a given threshold and such that the correlation value is at least as large as another specified threshold? In other words, is there a set $I' \subseteq I$ such that it simultaneously satisfies

$$\sum_{a \in I'} diff(a) \leqslant \theta \ , \qquad \sum_{a \in I'} corr(a) \geqslant \phi ?$$

Think of the *diff* and *corr* functions as encoding aspects of the target sought, but the recognition system is not permitted or is unable to recover the target explicitly by unraveling the *diff* and *corr* functions in order to use the information to help guide search. For example, consider the situation where you are running off to work, and you know you have forgotten something at home, but have no idea what or where it is. You start searching everywhere, for objects of all sizes and shapes, in all locations. The aha! you experience when you find the forgotten, unknown item is roughly analogous to the satisfaction of the *diff* and *corr* functions above. Even so, the unboundedness of the search is not quite complete, since you do not group totally arbitrary collections of image fragments as hypotheses for the unknown object. The unbounded search task is like trying to

make sense of the well-known Dalmation Sniffing at Leaves image using the strategy of grouping arbitrary blobs together until structure is seen.

Theorem 1. *Unbounded passive visual search is NP-complete*

The proof is by reduction from Knapsack and is given in [43]. It is well known that Knapsack has a pseudo-polynomial solution because it is a number problem, and in fact several logarithmic time, parallel solutions. These solutions yield approximate, not optimal, solutions within a given error bound and the time complexity varies with the setting of the error. However, it is argued in [45, 46] that none satisfy the constraints of biological plausibility (and the proposals of Brooks, Ramachandran and Ballard are intended to be biologically plausible). In any case, the methods by which such efficient solutions are found include hierarchical divide-and-conquer techniques and the use of intermediate representations, both of which are not permitted in the context of behaviorism. Moreover, there are other vision problems that are not formulated as number problems which are NP-complete and do not have such easily found approximation solutions (for example, [11, 25]). The use of a different implementation strategy such as neural networks does not help; the complexity class is for the problem, not the algorithm or implementation (recall that problem complexity is independent of algorithm or implementation). Also see [23] for a discussion of complexity and a set of theorems regarding the difficulty of learning procedures for neural networks.

The issue of approximate versus optimal solutions is an important one with respect to biological intelligence. It is probably true that for many tasks, an approximate solution suffices. In fact, this is a key point in the proposal by Tsotsos [44, 45] where a set of optimizations and approximations were presented in an attempt to achieve a biologically plausible architecture for vision. However, even to achieve this, certain compromises on processing must be made; the common ones, and those made by the Tsotsos proposal, are disallowed by the behaviorist dogma (hierarchical abstraction, intermediate shared representations, explicit targets, attentive processing, and so on). The search space is much too large to even find a set of acceptable yet sub-optimal solutions efficiently without specific combinatorics-defeating strategies.

2.1.2. Bounded passive visual search (BPVS)

The *diff* and *corr* functions here are based in the target data; that is, the test image subset is still sought as the solution, but the contents of the target image direct the search within the test image. Given a test image I, a target image T, a difference function, and a correlation function, is there a subset of pixels I' of the test image such that the difference between that subset and the corresponding subset of pixels in the target image is less than a given threshold and such that the correlation between the two is at least as large as another specified threshold? In other words, is there a set $I' \subseteq I$ such that it simultaneously satisfies

$$\sum_{t \in T} diff(t) \leq \theta \ , \qquad \sum_{t \in T} corr(t) \geq \phi \ ?$$

Theorem 2 (Tsotsos [43]). *Bounded passive visual search has linear time complexity in the number of test image locations.*

2.1.3. Unbounded active visual search (UAVS)

In [40], a view of active perception as an incremental hypothesis-test strategy was proposed. This strategy was joined with the visual search definitions presented earlier to arrive at a new formulation of the active visual search task. A necessary component is an input image sequence. The incremental strategy solves the visual search task for each image in the sequence, but instead of stopping the search process after the first satisfactory solution is found, it continues until all are found for each image. This forms the hypothesis set which is carried over to the next time instant. The thresholds on the *diff* and *corr* functions are tightened to constrain the solutions further and the process repeats. This incremental tightening of constraints may be viewed as passing the initial hypothesis set through an ever-narrowing sieve. The method is further elaborated in theory in [48], and in practice in [51].

The incremental strategy can be formalized as follows. Given a test image sequence in time I_t, $t = 1, \ldots, \tau$, a difference function, and a correlation function, is there a sequence of sets $\Im_t$ for $t = 1, \ldots, \tau$, where $\Im_t$ is the union of all sets $I'_t \subseteq I_t$, such that each I'_t satisfies

$$\sum_{a \in I'_t} diff(a) \leq \theta_t \ , \qquad \sum_{a \in I'_t} corr(a) \geq \phi_t \ ,$$

where

$$\theta_1 \geq \theta_2 \geq \cdots \geq \theta_\tau \ , \qquad \phi_1 \leq \phi_2 \leq \cdots \leq \phi_\tau \ .$$

Theorem 3 (Tsotsos [48]). *Unbounded active visual search is NP-complete.*

2.1.4. Bounded active visual search (BAVS)

The incremental strategy briefly described above applies equally well to the bounded version of the problem. Given a test image sequence in time I_t, $t = 1, \ldots, \tau$, a target image in time T_t, a difference function, and a correlation function, is there a sequence of sets $\Im_t$ for $t = 1, \ldots, \tau$, where $\Im_t$ is the union of all sets $I'_t \subseteq I_t$, such that each I'_t satisfies

$$\sum_{a \in T_t} diff(a) \leq \theta_t \ , \qquad \sum_{a \in T_t} corr(a) \geq \phi_t$$

where

$$\theta_1 \geq \theta_2 \geq \cdots \geq \theta_\tau \ , \qquad \phi_1 \leq \phi_2 \leq \cdots \leq \phi_\tau \ .$$

Theorem 4 (Tsotsos [48]). *Bounded active visual search has linear time complexity in the number of test image locations.*

Summarizing, if no targets are available, search performance may require at worst exponential time in the size of the image. If targets are known, and permitted to participate in top-down guidance, then performance is no worse than polynomial (approximately linear) in the same variable. These theorems provide strong evidence that general-purpose, data-directed vision is not only computationally intractable, but also biologically implausible [44, 45].

3. Perception within behaviorism

The last section argued for visual search as an integral part of intelligent behavior and showed that data-directed solutions to visual search are not acceptable in any theory of behavior. This section focuses on a number of related major points. First, arguments are presented as to why perception is not as easy as implied by the behaviorists and why all perceptual actions are not necessarily directly linked to action, i.e., are not externally observable. Then, several of the key principles of behaviorism will be tested against current biological knowledge, and it will be clear that those principles have no support.

3.1. Perception is not easy

Behaviorists imply that perception is easy: a set of primitives exist that can be used as input to behavior modules and that these primitives are next to trivial to compute. Only if this is true could they claim that the world be used as memory and visual information can be rapidly re-acquired on demand. It seems that such conclusions can only be obtained from an over-simplification of results from behavioral psychophysics.

There is much evidence from human behavior on visual search tasks for the "pop-out" effect: for certain visual displays, subjects can find targets in time independent of the number of distracting items in the display (see [39] for example). This response time is on the order of a few hundred milliseconds. This seems to be the source of the belief that perception can be easy. Those targets, however, are very specific (for example, a red target among a field of green ones, or a square in a field of lines, or the like). Moreover, in many of the experiments the target must be known explicitly in advance; in other experiments, the task is figure-ground segregation. Although there is a good deal of controversy as to how complex the differences between targets and distracters can be in order to ensure this behavior,[10] it is quite clear that many types of visual search behavior require time linear (or worse) in the number of items in the display (see [39, 41, 44, 45]).

[10] Some experimenters found that more complex target types, which are conjunctions of two or three features can also be found in constant time (Wolfe et al. [52], Nakayama and Silverman [32]). Others claim this is artifactual or that performance depends on the degree of dissimilarity between target and non-target (Tsotsos [45], Duncan and Humphreys [15]).

Even if the general structure of the display points to a linear response time curve, the slope of that curve varies significantly and is determined by the spatial arrangement and distribution of features, and the Y-axis intercept is determined by the "cognitive and motor overhead" of processing and responding to the display (this is not understood in general). Finally, the types of displays used are quite impoverished compared to natural images; they typically include no more than about 20 items, nicely segregated on a white background. In general, the characterization of which stimuli lead to what kinds of performance is not completely understood. It is important to note that there are really four different kinds of tasks that must be addressed: detection (is a target present in the test image or not?), localization (where is the target?); segmentation (which subsets of pixels represent it?), recognition (label the found subset of pixels with a semantically relevant entity). Of these, experimentalists have concentrated on detection, localization and recognition. All are of importance in a behaviorist context.

The bounded visual search problem described earlier predicts linear behavior for visual search where the target is known. If the mechanism of attentional control is added using knowledge of the target, constant time performance is predicted for pop-out displays [44, 45]. Therefore, if perception is to be fast the visual world as well as the intelligent agent must have special characteristics. The following conclusions are due to converging behavioral[11] and complexity analysis results.

(I) If the target (and position) is known, and it is distinguished from all distractors in an obvious manner[12] (colour for example), then detection, localization and recognition can be done in time independent of the number of items in the display [38].

(II) If the target (but not position) is known, and it is distinguished from all distractors in an obvious manner, then detection, localization and recognition can be done in time independent of the number of items in the display, but is slightly slower than case I [38].

(III) If the target (but not position) is known, and it is not distinguished uniquely in a simple way[13], then detection may require up to linear time in the number of display items [40].

(IV) If the target (and position) is known, and it is not distinguished uniquely in a simple way, then detection may require up to linear time in the number of display items, but is much faster than in type III [38].

(V) If the target is not known, but is distinguished in an obvious manner, then detection and localization may be done in time linear in the number of display items (for example, texture pop-out [4]).

---

[11] The citations included below only scratch the surface of the literature which is relevant. The reader is referred to the review article by Green [17].

[12] That is, it is a disjunctive type of display [38].

[13] That is, it is a conjunctive type of display [38].

(VI) Otherwise, detection and localization may require anything from a high slope linear function to polynomial to exponential time (with respect to the image size and/or feature space) [41, 44, 45]).

Detection and localization seem to always go together at least for the type I and II cases above [22]. Localization is a prerequisite for accurate recognition [31]. These six classes presented are by no means the complete set of classes. For example, in all of the above, single targets are assumed: what happens if there is more than one target? Although this account may present a somewhat clear version of events, the truth is that things get complicated very quickly if variations in experimental settings, features, etc. are considered. See [17] for a review.

The bounded visual search problem referred to in this paper has obvious counterparts (types I, II, III, and IV); the unbounded problem is associated with types V and VI. However a little care is required in making this distinction. It does not appear to be the case that arbitrarily complex targets can be used to guide search. Thus, there may be many situations when, even though the target is explicitly known, it is of such a form as to be unusable by our visual systems as guidance. For example, it is not likely true that arbitrarily complex colour combinations can lead to constant time target recognition in humans.[14] These search tasks would be included in the unbounded category.

Visual search behavior and theoretical evidence do not support the behaviorist position that vision can be fast except when targets in disjunctive displays are known explicitly or if target and location are known in conjunctive displays.

3.2. *All perceptual acts are not externally observable*

In Connell's incarnation of behavior-based architecture, a number of independent modules are proposed, each of which uses some type of perceptual information as input and produces commands for a creature's actuators; competing commands are merged with a hard-wired priority scheme. A module is equivalent to a production rule, perhaps with a small amount of state. Each module thus implements a small piece of the creature's overall behavior at the actuator level. Behavior is always observable and perception is always directly linked to action. This is clear in Brooks' use of the intelligence principle, namely, that intelligence is in the eye of the beholder.

In biological systems, all visual behavior does not lead to externally observable behavior and there is a clear example of this. It has been known for some time now that visual search behavior in humans is due in part to a form of attention

[14] Simply consider the various picture matching puzzles designed for children. Even though sometimes target images are provided, they seem not to help with target localization; detailed exhaustive search is still required. A particularly good example is the set of colour pictures in the "Where's Waldo?" series of books [20], which depict large numbers of soldiers engaged in battle. Even if the exact image of a target soldier is given, the search task does not seem to benefit from this knowledge in any useful manner. Part of the problem is that in these kinds of images, individual items are not segmented from one another (not separated with intervening background as in typical visual search experiments). Thus, one may hypothesize that the task of segmentation may be a major barrier to fast response.

termed "covert" [34]. Typical experimental paradigms for visual search require that the eyes remain fixated and thus there is no external or "overt" attention manifested by eye movements. The observed linear time behavior mentioned in the previous section is due to internal serial processes. Recall the definition of "scalability" presented earlier—that performance be the same in quality and timing. This aspect of timing is not captured by Brooks' subsumption scheme in the same formalism as is an external behavior.[15] In an important sense, this internal processing may be considered as a type of reasoning.

3.3. *Biological realities*

Most of the published behaviorist positions do not receive support from current knowledge of the neuroanatomy and neurophysiology of primate and human visual cortex. It is not possible to provide a thorough review of this material here and the reader is referred to Felleman and Van Essen's review [16] of the macaque visual cortex, the most complete collection to date of anatomical and physiological data on the functional organization of the monkey visual cortex.

Three behaviorists' claims go against biology: there are no intermediate representations; there are no hierarchical computations; and there is no explicit representation of goals. These will be addressed in turn.

One of the more established characteristics of the visual cortex is that it contains visual areas or maps. The current total of such areas for the macaque monkey is 32; no doubt all have not yet been discovered. These areas have been studied by anatomic methods, single-cell recordings, PET scanning and by lesion studies over the past twenty years. Although the story is far from complete, and the exact function of each area is not fully understood, it is clear that each area represents the result of some processing of visual input; the sensitivities and selectivities of the neurons within the areas are one of the key determinants of area segregation. The representation may be considered to be the array of firing rates of the population of neurons in a given cortical area. Many areas seem to represent the whole visual field; others, only parts of it.

An interesting connectivity pattern has emerged among the 32 areas. The total number of inter-area connections discovered so far totals 305. Most of the 305 connections form reciprocal pairs (121 pairs). Of the reciprocal pairs, 65 pairs have been clearly identified as ascending/descending pairs. Second, the areas are organized into a hierarchy of 14 layers deep, beginning with the retina at the input and ending with the hippocampus at the top. Ten of those layers are visual processing layers within the visual cortex. There are no connections that span the full 14 layers; the greatest span is 7 layers and the majority of connections span only 1 or 2 layers. The connectivity figures do not include the connections to lower brain areas; most of the cortical areas have connections to lower visual and oculomotor areas (superior colliculus, thalamic reticular nucleus, pulvinar com-

[15] This is only one type of internal behavior out of possibly a large number. Are humans who are sensory-deprived not intelligent?

plex, basal ganglia). Information from the retina also goes directly to the superior colliculus. A role for such connections in eye movement behavior is hypothesized. Only 4 of the 32 areas have connections (so far) to the somatosensory and motor areas of the brain; these areas are all within the top few layers of the hierarchy. These 4 areas receive connections from only 8 other visual areas in total. Information from the retina necessarily makes many stops along the way, being processed by several areas (at least including the lateral geniculate nucleus and V1 and 14 other intermediately positioned areas in 6 different layers of the hierarchy if the shortest paths to the "output" areas are considered; see the figures in [16]). It is clear that the visual cortex uses both hierarchical organization as well as intermediate representations.

It is becoming clear that task knowledge has a great effect on visual processing. Individual neurons (in monkey) in areas considered early in the processing sequence (V1, V2 and V4, [19, 29, 30]) have been observed to change their tuning properties as a result of task knowledge. Some neurons in V4 even appear to code the cue provided to the subject for a particular experiment (target orientation in an orientation selection task). Even more surprising is the observation that the orientation cue may be provided through tactile stimulation rather than visual, and then is stored in area V4 neurons of the visual cortex [18]. Knowledge of task, that is the goal, not only has great effect on processing, but seems to be represented explicitly within the processing hierarchy. Indeed, in area IT, cues seem to not only be represented but the representation is retained for several seconds in anticipation of a task [9].

Although many have hoped that simplifying assumptions may be made in order to model the human visual system computationally, the evidence is pointing away from this. For example, the concept of independent modules received a great deal of attention within the computational vision community for some time and behaviorists have embraced this notion as well. There is no reason to believe that any of the visual areas process information independently of one another; the connectivity patterns do not support this. Lesion studies demonstrate most clearly that if one area of the brain is destroyed or a connection is cut, the remaining areas at the individual neuron level as well as at a gross functional level yield different functions (clinically, this occurs naturally in humans as a result of a stroke). Another concept which appeared to permit simplifications was the segregation of "what" and "where" processes into independent pathways. This too has very recently fallen into doubt; higher visual functions leading to the perception of spatiotemporal relationships or to visual object recognition do not depend exclusively on information processed by either pathway [28]. Finally, the fact that a given neuron in the visual hierarchy, even as early as area V1, can change its functional properties as a result of task (that is, external) influences is all the proof that is required to put the independent modules concept to rest. The notion that vision may be modeled by integrating independent modules is not viable.

The description above focuses on the visual cortex. Since the behaviorist approach relies on fast perception and the ability to find stimuli that trigger

behaviors as integral components, its principles must also extend to perceptual processing. There is little biological support for behaviorist principles as they apply to perception.

4. Behaviorism and the scaling problem

There are four major components to a behaviorist solution (call the problem "stimulus–behavior search"):

- localize and recognize a stimulus (with no pre-specified target);
- link the stimulus to an applicable action;
- decide among all applicable actions,
- generate actuator commands.

The first subsection will present two theorems relating to the problem complexity of stimulus–behavior search

4.1. Complexity of stimulus–behavior search

Since the unbounded passive visual search problem (UPVS) described in Section 2 is exactly the first of the four components of the behaviorist solution given at the beginning of Section 4, the following theorem is stated without proof:

Theorem 5. *Unbounded passive stimulus–behavior search is NP-hard.*

It must be stressed that the NP-hardness, like for unbounded passive visual search is due entirely to the combinatorics of searching an image without an explicit target. The fact that there is additional processing for determining the applicability of behaviors and for deciding which behavior to execute only makes the time complexity of the problem worse than that of unbounded visual search.

Will active behaviorism help? The addition of the time dimension so that the search for a behavior–stimulus pair is over time does not necessarily help. In [47, 48] it was shown that if two problems can be solved by both active and passive methods, then the active approach will be more efficient only under certain constraints having to do with the amount of memory available for storage of intermediate results, the extent of sensor movements, the size of the visual field and so on. More importantly, an intermediate representation of best hypotheses is required in order to satisfy those constraints (and is implied in Bajcsy's original paper [2]). The use of such intermediate representations is not within the behaviorist paradigm. In any case, the following theorem follows directly from Theorem 3 and the task decomposition given at the beginning of this section:

Theorem 6. *The unbounded active stimulus–behavior search problem is NP-hard.*

As described in Sections 2.1.2 and 2.1.4, the bounded visual search problem can be linear with respect to image size. Although no argument supporting the

following claim is given here, it is likely that the bounded stimulus–behavior search problem has low order time complexity also.

4.2. Implications

Recall that the central thesis of this paper is to argue against the behaviorist claims that the theory will scale up to problems which are human-like in their size and, that behaviorism can be used as a model of human intelligence. Much discussion has already been presented with respect to how poorly the theory fits with biological evidence. The previous section adds to this evidence the fact that given our current understanding of computational complexity, the behaviorist philosophy presents a view of human intelligence as a computationally intractable activity. Since human intelligence is an existence proof for its own tractability, in fact, this means that the behaviorists are defining and then attempting to solve the wrong problem!

A couple of small modifications to the behaviorist position have appeared in implementations which are claimed to make a difference; they do not. The addition of state to behaviors does not affect this result at all since the NP-hardness is due to the perception component alone. The use of many special small sensors instead does not change the NP-hardness. Moreover, humans do not have many special visual sensors, they have only two eyes. These points are further discussed in the next section.

4.3. A numerical exercise

The behaviorist position, if relaxed, might in fact result in biologically plausible solutions; these modifications to the theory would be anathema to its proponents. This section presents a little numerical exercise whose goal it is to show how necessary the modifications are.

Above it was shown that bounded visual search may be fast, so it is natural to suggest that perhaps the behaviorist strategy could employ bounded search rather than unbounded search. Let us numerically check whether relaxation of the "no targets" principle is sufficient to permit scalability. Let us permit explicit representations of stimuli that trigger behaviors (the Herbert robot does exactly this: Herbert contains an explicit procedure for the recognition of soda cans, and can recognize them only at a certain height and in certain positions, [10]). Thus, all behaviors can run in parallel, each has complete knowledge of its trigger stimulus and performs perceptual processing independently of the other behaviors. Each behavior can extract a stimulus relatively quickly. Can this architecture now satisfy the scaling definition presented earlier?

It is required to not only specify the size of the input data as was done in the definition of Section 1, but also, to specify the number of possible responses to this data (or behaviors). How can this be done? One simple-minded (and perhaps not the best) way follows. The *Visual Dictionary* [12] contains over 3,000 images

portraying over 25,000 generic objects in all (cars, stadiums, forks, etc.). These images do not include:

- variations in object context which could lead to very different behaviors (picking up a fork at the dinner table while seated is very different from picking up the fork your child dropped behind the living room couch);
- variations due to object type or brand (the act of sitting on a chair is virtually identical to the act of sitting on a sofa even though the images are very different; there are scores of brands of cars, and we all would drive a sleek, red, 2-seat sports car very differently than we would drive the family sedan);
- variations due to colour (the human eye is sensitive to about 500 different hues, 20 steps of saturation for each and 500 levels of brightness for each hue-saturation pair);
- variations due to lighting (the eye is sensitive to luminance spans of 10,000,000,000 to one);
- variations due to viewpoint (seeing an object from one viewpoint may cause a shift to another viewpoint in order to fully recognize it; depending on the resolution of the sensing system and on the method of perceptual processing, up to 30% of the viewing sphere around a single object may contain degenerate viewpoints which require sensor motion for disambiguation, [50]);
- time-varying events.

Would 1,000,000 behaviors be sufficient? 100,000,000? It is very hard to say; however, it is clear that the number is necessarily very large for the average adult. Let us use 25,000,000 for the sake of argument. Thus, on average, each of the 25,000 generic objects in the dictionary:

- can be found in 4 different spatial contexts;
- can be seen under 4 different lighting conditions;
- can be of 2 different types;
- can have one of 4 distinct colours;
- can be seen from 4 distinct viewpoints.

For each of these objects, the most primitive of behavior sets is included. Each can elicit one of two behaviors: do something or ignore the object. This seems to be a very conservative estimate for the total number of human behaviors considering that only visual stimuli and their resulting behaviors are counted!

Given 2 retinas with about 250,000,000 photoreceptors, each perception module for each of the 25,000,000 behaviors must be directly connected to each photoreceptor (in order to ensure translation invariant recognition). The total number of connections would be 6.25×10^{15}. That is already larger than the total number of synapses in the human brain (and remember only visual behaviors are counted). As a model of intelligence, behaviorism with explicit targets is untenable.

Suppose that some machine someday can easily handle that connectivity. The model now requires 25,000,000 independent processing pathways, each beginning with targeted analysis of the visual image. Do any of the 25,000,000 analyses have any processing steps in common? Of course they do; it is hard to imagine any perception process that would not use methods to detect image structure (for

example, discontinuity, homogeneity, temporal variation) or if not, to use some kind of matching procedure on the data directly that would ensure colour constant, translation and rotation invariant, motion invariant perception of object identity. Would it not enhance the efficiency of processing resources and system power consumption if those steps were unified, performed once, and then the results shared? Of course, and thus is born the intermediate representation; behaviorism disallows such representations.

Suppose that hardware is so cheap and power so plentiful that the replication of representations (so that they are not shared) is not really a concern. The output of the 25,000,000 behaviors must now be processed such that the most appropriate behavior is the overall system response. A priority network is needed (as in Connell's implementation of the subsumption scheme) for deciding this. In the subsumption scheme, each behavior may spy on and inject signals into any other lower level behavior. Assume that there are L levels, each with an equal number of behaviors, and compute the total number of connections needed. Further assume that each behavior on average only affects $1/n$th of all the behaviors in the levels below it. Thus, each of the top level's $25{,}000{,}000/L$ behaviors are connected to $(L-1)25{,}000{,}000/(nL)$ behaviors; the next level down also has $25{,}000{,}000/L$ behaviors and each of them is connected to the $(L-2)25{,}000{,}000/(nL)$ behaviors below it and so on. How many connections are there in total in the priority net alone? Some simple calculations lead to the total of

$$\sum_{x=1}^{L-1}\left(\frac{25{,}000{,}000}{L}\right)\left(\frac{L-x}{L}\right)\left(\frac{25{,}000{,}000}{n}\right) = 6.25\times 10^{14}\,\frac{(L-1)}{nL}$$

connections. Stated otherwise, there would be $B^2(L-1)/(nL)$ connections, where the number of behaviors is represented by B. To this number one must add the number of connections required at the sensory end as derived earlier; the total number of connections for the input plus the priority net is

$$6.25\times 10^{14}\times(10+(L-1)/(nL))\,.$$

Even if very sparse connectivity in the priority net is assumed (n is large), this total is much too large for a biologically plausible model of intelligent behavior. From a machine intelligence standpoint, very optimistic assumptions must be made.

The above calculations assumed static images; the world is a dynamic one and we react to motion and to moving objects and changing events. How many more behaviors must be added to account for the dynamic nature of the world? It is clear that it is not easy to estimate this figure and it is equally clear that it must be very large. In a sense, behaviorist implementations which permit each behavior to contain state account for some of the time-varying aspects of the world. However, following behaviorist principles, state is not permitted to affect the sensory processing; this would mean that processing would be guided or targeted in some sense. Exactly the same conclusions as above are reached insofar as the perception portion of processing is concerned. Since the limits were exceeded

without time-varying images, they are exceeded in a much more obvious manner once these are included in this numerical exercise.

It seems that regardless of the direction one takes in this exercise, some behaviorist principle must be violated or questioned. The difficulty of perception is the issue and not the behaviors themselves.

What about other modifications to the paradigm? Would many small, special-purpose sensors help? Let us leave aside for the moment that humans have two eyes and not a set of many small special visual sensors; this strategy seems to have no relation to human biology. Suppose that each behavior has its own sensor so that there are 25,000,000 sensors. Let their average number of receptors (or pixels) be given by R. R must be at least 10 in order to at least cover the visual field of the two eyes. Overall connectivity seems reduced to an acceptable degree, yet small values of R mean that each perception module is position-specific. In order to enable for translation-invariant recognition, all of the behaviors must have access to the output of all the sensors undoing the connectivity benefit and/or necessitating the representation of results which are shared by all behaviors.

The strict behaviorist position, even after permitting very optimistic hardware expectations, does not scale to human-like problems as defined in Section 1; it also does not scale in a biologically plausible manner. Note that this exercise was performed using the size of visual problems only; the remaining senses also require analysis. Relaxing some of the principles seems necessary; this is true especially as the scaling issue pertains to a model of human behavior. One may argue that the number of behaviors used in this "back of the envelope" calculation is wildly too large. It depends on what one means by behavior. Here, since behaviors must employ bounded visual search, and taking the lead from Connell's implementation (soda cans in one orientation, at one height and one distance from the robot), it is not at all difficult to imagine far more behaviors than 25,000,000 if every motor action that has slightly different parameters (speed, grasp strength, arm pose and trajectory in space, etc.) is considered as a unique behavior since it would be necessitated by a slightly different target image.

4.4. *Summary: scaling and computational behaviorism*

Several key points must be emphasized:

- *Visual search can be viewed within the behaviorist paradigm and is an integral component of any intelligent behavior.* Recall that for a visual search task, a subject is presented with a target (or targets), and a test image and asked to determine whether or not that target is present in a test image. This involves two behaviors (or stimulus–action pairs): (1) if target present, press button A; (2) if target absent, press button B.
- *Behaviorism with embedded unbounded visual search problems may require exponential time (exponential in image size) for the signal matching tasks.* If the target is given as a set of constraints or is known only implicitly in some way, behaviorism will not scale for realistic, nontrivial images.

- *Active behaviorism requires intermediate representations of hypotheses by definition.* Although the terms active vision, reactive behavior, animate vision etc. are used in the literature almost as synonyms, in fact none truly reflect active perception as defined by Bajcsy [2] simply because they do not acknowledge the need for intermediate representations of hypotheses [48].
- *Strict behaviorism cannot satisfy the definition of scaling in Section* 1. Data-directed, spatial parallelism simply does not make sense if human-size problems and behavior spaces are considered.
- *The strict behaviorist position must be modified in order to satisfy the scaling definition presented earlier.* The appropriate kind of scaling in the context of human visual behavior may be accomplished by permitting optimizations and approximations of the kind described in [44, 47, 48]: intermediate representations, task guidance, visual attention, hierarchical organization, spatial abstraction, logically segregated visual maps. This set has been shown to be sufficient (but not necessary) to ensure scalability as defined in Section 1 with human-like resources, problem sizes and performance specifications.

5. What do the behaviorists *really* mean?

The strict behaviorist position has too many problems; the claims of biological relevance and scaling simply are not supportable. But, is the strict position really what the behaviorists mean? The strict position is certainly the cause of controversy: the controversy is due to the use of slippery terms such as goals, hierarchies, representations, intermediate, perception, etc. What exactly do these mean in the context of behaviorism? Three examples will be given now that show that the arguments presented in print by the behaviorists and the implementations of their robots in practice do not agree with regard to these terms.

Targets and goals

In the previous sections, it was argued that perception with no targets is not necessarily fast. Ballard's claim that the external world can be used as a visual memory because its contents can be re-acquired quickly is not in most cases valid. Moreover, biological vision seems to make use of task information and perhaps to even represent targets explicitly. But, the term target as used in the definition of UPVS referred to a model of what is being sought such that it was sufficient to constrain search in image space and it was applied in a processing architecture able to optimize search given knowledge of the target. If the target is known (the target or goal is explicitly represented and/or realized by the circuit) the behaviorist paradigm can be very fast even if targets can be rotated and scaled (although this has not appeared in any of the behaviorist implementations). This corresponds exactly to the existing realizations of subsumption architecture. Targets are always specified. This kind of goal is not the same as the type of goal a planning system may be given (starting in room A go to room D), but it is a goal

nevertheless. The behaviorists, it seems, are arguing against the latter, but their position on the former is unclear.

Hierarchical computations

In print, Brooks uses the term hierarchy in the same sentence as subroutine calls [7]. It is highly doubtful that the brain does much in the way of calling subroutines in the same way conventional programming languages do. However, there seems little doubt that the brain does use hierarchical processing, one level providing input to the next, in both data-driven and knowledge-driven directions. Is this not exactly how the different levels of a subsumption architecture interact? The priority network which connects and arbitrates among behaviors in the subsumption scheme is an example of a hierarchical computation. Further, Connell [10] suggests the use of partitions for grouping behaviors and for levels of arbitration. In this scheme, a group of behaviors can be switched on and off extending the capability of his Herbert robot so that it might recognize more objects and could treat them differently. In fact, he suggests another good use of hierarchical processing.

Intermediate representations

In the brain, intermediate representations seem to be everywhere; even within visual areas, many separate populations of neurons have been found to encode the same visual space but in different ways. Yet the published strict behaviorist dogma claims there is no need for such representation. However, in the control system diagrams published by Brooks [5] and by Horswill and Brooks [21], examples of intermediate representations (using the definition presented earlier) can be found. In the former case, a sonar map is used as input for two behaviors (halt and move forward) while in the latter, a representation of object features (centroid and x, y coordinates) is used as input for two behaviors (drive and turn).

What do the computational behaviorists really mean?

6. Summary

Strict adherence to the philosophy of the behaviorists means that efforts to demonstrate that the paradigm scales to human-size problems are certain to fail, as are attempts to evaluate it as a model of human intelligence. The strict position thus cannot be what the behaviorists really mean. It would benefit the research community if they could elucidate their terms, and provide theoretical arguments that support claims of scalability. This is not to say that the research by those researchers is not useful. Far from it: the exercise has proved very important. For small behavior sets, and well understood and small sensor signals, the behaviorist solution seems to lead to useful devices. The combinatorial problems described above are avoided by ensuring that the number of "pixels" of signal is very small.

The behaviorist position will lead to successful special-purpose robots—this is not a small accomplishment. However, it is not the solution, as it is currently

stated, to intelligent behavior. The issue of scaling to human-like problems requires a much deeper analysis of the amount of computation required for human-sized problems plus much more serious consideration of actual human behavior and neurobiology. Superficial correspondences are misleading.

Acknowledgments

Sven Dickinson, Michael Black, Suzanne Stevenson and Michael Gruninger provided valuable comments, as did an anonymous reviewer. This research was funded by the Information Technology Research Center, one of the Province of Ontario Centers of Excellence, the Institute for Robotics and Intelligent Systems, a Network of Centers of Excellence of the Government of Canada, and the Natural Sciences and Engineering Research Council of Canada.

References

[1] Y. Aloimonos and A. Rosenfeld, Computer vision, Science **253** (13 September 1991) 1249–1254.
[2] R. Bajcsy, Active perception vs passive perception, in: *Proceedings IEEE Workshop on Computer Vision: Representation and Control*, Bellaire, MI (1985) 55–62.
[3] D. Ballard, Animate vision, *Artif. Intell.* **48** (1991) 57–86.
[4] J. Bergen and B. Julesz, Rapid Discrimination of Visual Patterns, *IEEE Trans. Syst. Man Cybern.* **13** (1983) 857–863.
[5] R. Brooks, A layered intelligent control system for a mobile robot, *IEEE J. Rob. Automation* **2** (1986) 14–23.
[6] R. Brooks, Intelligence without representation, *Artif. Intell.* **47** (1991) 139–159.
[7] R. Brooks, Intelligence without reason, in: *Proceedings IJCAI-91*, Sydney, Australia (1991) 569–595.
[8] R. Brooks, New approaches to robotics, *Science* **253** (13 September 1991) 1227–1232.
[9] L. Chelazzi, E. Miller, J. Duncan and R. Desimone, A neural basis for visual search in inferior temporal cortex, *Nature* **363** (1993) 345–347.
[10] J. Connell, A colony architecture for an artificial creature, Ph.D. Thesis, AI-TR1151, MIT AI Lab., Cambridge, MA (1989).
[11] M. Cooper, *Visual Occlusion and the Interpretation of Ambiguous Pictures* (Ellis Horwood, Chicester, England, 1992).
[12] J. Corbeil, *Visual Dictionary* (Stoddart, Toronto, Ont., 1986).
[13] H. Davson, *The Eye: Visual Function in Man, Vol. 2A* (Academic Press, New York, 1976).
[14] J. Dowling, *The Retina: An Approachable Part of the Brain* (Harvard University Press, Cambridge, MA, 1987).
[15] J. Duncan and G. Humphreys, Visual search and stimulus similarity, *Psychol. Rev.* **96** (3) (1989) 433–458.
[16] D. Felleman and D. Van Essen, Distributed hierarchical processing in primate cerebral cortex, *Cerebral Cortex* **1** (1) (1991) 1–47.
[17] M. Green, Visual search, visual streams and visual architectures, *Perception & Psychophys.* **50** (4) (1991) 388–403.
[18] P. Haenny, J. Maunsell and P. Schiller, State dependent activity in monkey visual cortex II. Retinal and extraretinal factors in V4, *Experimental Brain Res.* **69** (1988) 245–259.
[19] P. Haenny and P. Schiller, State dependent activity in money visual cortex I. Single cell activity in V1 and V4 on visual tasks, *Experimental Brain Res.* **69** (1988) 225–244.

[20] M. Handford, *Where's Waldo?* (Grolier Ltd., Toronto, Ont., 1987).
[21] I. Horswill and R. Brooks, Situated vision in a dynamic world: chasing objects, in: *Proceedings AAAI-88*, St. Paul, MN (1988) 796–800.
[22] J. Johnston and H. Pashler, Close binding of identity and location in visual feature perception, *J. Experimental Psychology: Human Perception and Performance* **16** (4) (1990) 843–856.
[23] S. Judd, *Neural Network Design and the Complexity of Learning* (MIT Press, Cambridge, MA, 1990).
[24] E. Kandel and L. Squire, Cognitive neuroscience, *Current Biology* **2** (2) (1992) 143–145.
[25] L. Kirousis and C. Papadimitriou, The complexity of recognizing polyhedral scences, *J. Comput. Syst. Sci.* **37** (1988) 14–38.
[26] D. Kirsh, Today the earwig, tomorrow man?, *Artif. Intell.* **47** (1991) 161–184.
[27] M. Mataric, Perceptual parallelism and action selection as alternatives to selective perception, in: *Working Notes, AAAI Symposium on Control of Selective Perception*, Stanford, CA (1992) 96–99.
[28] J. Maunsell, Functional visual streams, *Current Biology* **2** (4) (1992) 506–510.
[29] J. Moran and R. Desimone, Selective attention gates visual processing in the extrastriate cortex, *Science* **229** (1985) 782–784.
[30] B. Motter, Focal attention produces spatially selective processing in visual cortical areas V1, V2 and V4 in the presence of competing stimuli, *J. Neurophysiol.* **70** (3) (1993) 909–919.
[31] J. Müller and P. Rabbitt, Spatial cueing and the relation between the accuracy of where and what decisions in visual search, *Q. J. Experimental Psychol.* **41A** (4) 747–773.
[32] K. Nakayama and G. Silverman, Serial and parallel processing of visual feature conjunctions, *Nature* **320** (6059) (1986) 264–265.
[33] U. Neisser, *Cognitive Psychology* (Appleton-Century-Crofts, New York, 1967).
[34] M. Posner, Orienting of attention, *Q. J. Experimental Psychol.* **32** (1980) 3–25.
[35] P. Rabbitt, Sorting, categorization and visual search, in: E. Carterette and M. Freidman, eds., *The Handbook of Perception: Perceptual Processing, Vol. IX* (Academic Press, New York, 1978).
[36] V.S. Ramachandran, Interactions between motion, depth, color, and form: the utilitarian theory of perception, in: C. Blakemore, ed., *Vision: Coding and Efficiency* (Cambridge University Press, New York, 1990) 346–360.
[37] *SIGART Bulletin* **2** (4) (1991) 12–184; Special Section of Integrated Cognitive Architectures.
[38] A. Treisman, Preattentive processing in vision, *Comput. Vis. Graph. Image Process.* **31** (1985) 156–177.
[39] A. Treisman, Features and objects: the fourteenth Bartlett memorial lecture, *Q. J. Experimental Psychol.* **40A** (2) (1988) 201–237.
[40] A. Treisman and G. Gelade, A feature-integration theory of attention, *Cogn. Sci.* **12** (1980) 99–136.
[41] A. Treisman, and S. Sato, Conjunction search revised, *J. Experimental Psychology: Human Perception and Performance* **16** (1990).
[42] J.K. Tsotsos, A 'complexity level' analysis of immediate vision, *Int. J. Comput. Vis.* **1** (4) (1988) 303–320.
[43] J.K. Tsotsos, The complexity of perceptual search tasks, in: *Proceedings IJCAI-89*, Detroit, MI (1989) 1571–1577.
[44] J.K. Tsotsos, A complexity level analysis of vision, *Behav. Brain Sci.* **13** (3) (1990) 423–455.
[45] J.K. Tsotsos, A little complexity analysis goes a long way, *Behav. Brain Sci.* **13** (3) (1990) 459–469.
[46] J.K. Tsotsos, Is complexity analysis appropriate for analyzing biological systems?, *Behav. Brain Sci.* **14** (4) (1991) 770–773.
[47] J.K. Tsotsos, Image understanding, in: S. Shapiro, ed., *Encyclopedia of Artificial Intelligence* (Wiley, New York, 1992) 631–663.
[48] J.K. Tsotsos, On the relative complexity of active vs passive visual search, *Int. J. Comput. Vis.* **7** (2) (1992) 127–141.
[49] J.B. Watson, *Psychology from the Standpoint of a Behaviorist* (Philadelphia, 1919).

[50] D. Wilkes, S. Dickinson and J.K. Tsotsos, Quantitative modelling of view degeneracy, in: *Proceedings Eight Scandinavian Conference on Image Analysis*, Tromso, Norway (1993).

[51] D. Wilkes and J.K. Tsotsos, Active object recognition, in: *Proceedings CVPR-92*, Urbana, IL (1992) 136–141.

[52] J. Wolfe, K. Cave and S. Franzel, Guided search: an alternative to the feature integration model for visual search, *J. Experimental Psychology: Human Perception and Performance* **15** (1989) 419–433.