The hierarchical evolution in human vision modeling

Dana H. Ballard and Ruohan Zhang
Department of Computer Science, The University of Texas at Austin
Austin, TX, 78712, USA
danab@utexas.edu, zharu@utexas.edu

Abstract

Computational models of primate vision took a major advance with David Marr's tripartite separation of the vision enterprise into the problem formulation, algorithm and neural implementation, however many subsequent parallel developments in robotics and modeling greatly refined the algorithm descriptions into very distinct levels that complement each other. This review traces the time course of these developments and shows how the current perspective evolved to have its own internal hierarchical organization.

Introduction

The computational approach of the brain's realization vision dates from Marr's seminal theorizing that triaged the effort into problem specification, algorithm, and biological implementation. This organization decoupled the functional model of experimental observations from a complete account of all its neural details, which were extraordinarily complex. However, in the forty years since, the algorithmic approach to vision modeling has undergone elaborations into separate levels of abstraction. One result is a triage into three distinct levels [Ballard, 2015]. A *neural level* encompasses Marr's original functional level. Its models primarily respect low-level abstractions of the visual cortical anatomy. An *embodiment level* recognizes the embodiment of vision in an active agent that utilizes visuomotor behaviors. Models at this level include image acquisition and actions via an abstract motor system. The use of embodiment here is defined by the grounding of the body, as emphasized by many authors [Ballard et al., 1997, Noë, 2009, Clark, 2008]. An *awareness level* models the behavior by an explicit conscious agent. Models at this level include instructions comprehension of audio instruction and planning [Graziano, 2013, Dehaene, 2014].

While all three levels ultimately have a neural realization, the last two are distinguished by having less and less dependence on neural descriptions, analogous to the way silicon computers use separate computational levels. Thus embodiment models feature the dynamics of physical systems, and awareness models feature abstract symbolical descriptions. This paper reviews these developments, situating them on the historical timeline.

Level	**Description**	**Timescale**
Neural	Dynamics of neural circuits	$20 \sim 300$ milliseconds
Embodiment	Primitive behaviors defined by fast sensori-motor coordination such as a fixation and pickup of an object	$300 \sim 1000$ milliseconds
Awareness	Use of simulation to modify behavioral descriptions	10 seconds

Table 1: Functional levels for the study of neuroscience of behavior.

I. The functional neural level: The Marr paradigm

The neural investigation of human vision saw a huge advance with Hubel and Wiesel's studies of edge orientation retinotopic maps in striate cortex [Hubel and Wiesel, 1962]. Retinotopic maps of other basic image features such as color, velocity, direction selectivity, and stereo were identified soon after. But how could these features be computed? The answer awaited Marr and Poggio's seminal paper [Marr and Poggio, 1976] to introduce the advantage of pursuing a computational account of random dot stereo to introduce a functional approach. The calculation had to be done at least in the striate cortex as it was the first map that had cells with binocular receptive fields. The elegant algorithm maximized binocular matches that had similar depths. Its success ushered in the acceptance of an abstract algorithmic approach to account for biological observations.

Besides stereo, striate cortex recordings revealed several retinotopic maps of properties such as optic flow, color, and shape from shading that were characterized as a "two and a half dimensional sketch" [Marr, 1982] and "intrinsic images" [Barrow et al., 1978], which seemed to have common abstract functional constraints. These retinotopic two dimensional property images turned out to have two parameters per property, but only one fundamental equation relating the image's grey-level pixels to the parameters. For example, a retinotopic optic flow image measures the local x and y velocity vectors at each point in an image [Horn and Schunck, 1981]. The basic equation was that for every point in an image $I(x, y, t)$ the total variation had to be zero, i. e. $\frac{d(I(u,v,t)}{dt} = 0$, which could be expanded to

$$\frac{\partial I}{\partial x}u + \frac{\partial I}{\partial y}v = -\frac{\partial I}{\partial t}$$

Thus at each point in the image, there was the above equation with two unknowns, $u(x, y, t), v(x, y, t)$. To make the system of equations well-posed, typically spatial smoothness, that the changes in u and v would vary smoothly across the image was invoked to provide the needed additional constraints.

The impact of these results cannot be overestimated. Observable properties of cells in early cortex maps were understandable in terms of local constraints that potentially could be compatible by local networks.

It was time to introduce the image's observer. What constraints would drag one's attention to one item in an image over another? The proposed answer was hugely influenced by the work of Triesman [Treisman and Gelade, 1980] whose work suggested that combination local image features might be an attention stimulus. Objects defined with simple features such as color "popped out." Thus a green object was quickly located in a cloud of red objects, whereas combinations of features seemed to require that objects to be inspected individually, taking much more time. Thus the concept of image saliency was born with the seminal paper by Koch and Ullman [1987]. While with the hindsight of much later research shows that most situations require knowing the observer's goals to account for looking patterns [Triesch et al., 2003, Tatler et al., 2011], in many cases saliency provides an informative choice of useful image features [Koch and Ullman, 1987, Itti and Koch, 2001].

Like saliency, ever exacting Primal Sketch models made creative assumptions to achieve their results, but it burned out that at the theorizing Marr level, their assumptions have had to be revised. Most of all was that the smoothness constraint was challenging for natural images. Such images were rife with discontinuities that could defeat that objective function guiding the recovery of the properties from Marr's two and a half degree sketch (two spatial coordinates and a physical property).

Another issue was the models' locality. The decision to focus on the computations that could be modeled locally within the striate cortex retinotopic map. While it was known from the anatomy

that there were connections from more abstract cortical maps, Van Essen et al.'s definite hierarchical anatomy maps were not published until 1992 [Van Essen et al., 1992]. Thus, a research strategy was to constrain the models to have locally striate cortex connections.

Finally, a general assumption of the time was to emphasize two-dimensional image inputs. This stance was widely held by the general vision research community at the time. The thinking was that two-dimensional images were simpler than three-dimensional images and thus should be a priority. After all, the animal data could be obtained from anesthetized preparations, and the additional complication of an awake behaving animal posed new technical difficulties. The research impetus was to get the issues with two-dimensional images solved first and then move on to the more complicated time-varying three dimensions of the natural world.

II. The embodiment level: Active vision

Bajcsy was the first to recognize that seeing was an active process and that the computer vision enterprise had to move on [Bajcsy, 1988]:

```
"Most past and present work in machine perception has involved extensive
static analysis of passively sampled data. However, it should be axiomatic
that perception is not passive but active. Perceptual activity is exploratory,
probing, searching; percepts do not simply fall onto sensors as rain
falls onto ground."
```

This quote has at least two new ideas. One is that perceptions are not passive, but are ordered up by an active process of the perceiver. Another is that rather than a complication, active perception provides additional constraints for the perceiver. Thus the ability to move the head provides a straightforward relationship between depth and optic flow. Other depth cues from binocular vergence movements could program the eye's rotational vestibulo-ocular reflex appropriately.

The Bajcsy laboratory's binocular computer-controlled moving camera system shown in Fig. 1 was an instant success when it was unveiled at the IEEE conference in Traverse City, Michigan, in 1985. Thus in one instant, the computational vision enterprise became embedded into the world of the animate perceiver where the vision properties of the cortex had to deal with a world regularly in motion. Bajcsy had added another perspective to the Marr paradigm. Besides an algorithmic level confined to abstract neural circuits, a new level, including the perceiving agent's sensorimotor apparatus, was added.

The new level required new hardware. To access this world required moving cameras and sufficient computational cycles to keep up with the image video inputs, and quickly other designs emerged at Rochester [Brown, 1988] and KTH in Sweden [Pahlavan and Eklundh, 1992] had real-time processing. Rochester's binocular cameras were mounted on a platform that gave each camera independent yaw control, and that platform had independent pitch control. The system had the capability of making saccadic fixations at a considerable fraction of 700 degrees per second– the speed of human saccades, and could make vergence and pursuit movements. The real-time processing system, shown in Fig. 2, used $DataCube^{TM}$ video processing circuit boards. The boards contained a series of circuits, each containing programmable functionalities. A given control strategy typically used several such boards, which could be interconnected. For example, to compute kinetic depth, separate boards would compute vertical and horizontal image gradients. Knowing the fixation point allows inferring the sign of each point's depth, as points behind the fixation point move in a direction opposite to the camera motion and vice versa for points in front of the fixation point. Thus a large subset of image processing could take advantage of the new constraint

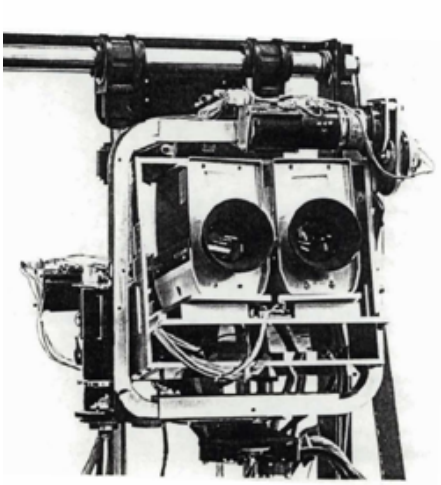

Figure 1: **Bajcsy's camera system.** A joint project between the University of Pennsylvania's computer science and mechanical engineering departments, It was the first to combine pitch and yaw degrees of freedom in a binocular system. The known movements provided egocentric parametric information that simplified many visual computations.

of self-motion to use much simpler and more robust algorithms to recover physical properties from image pixels.

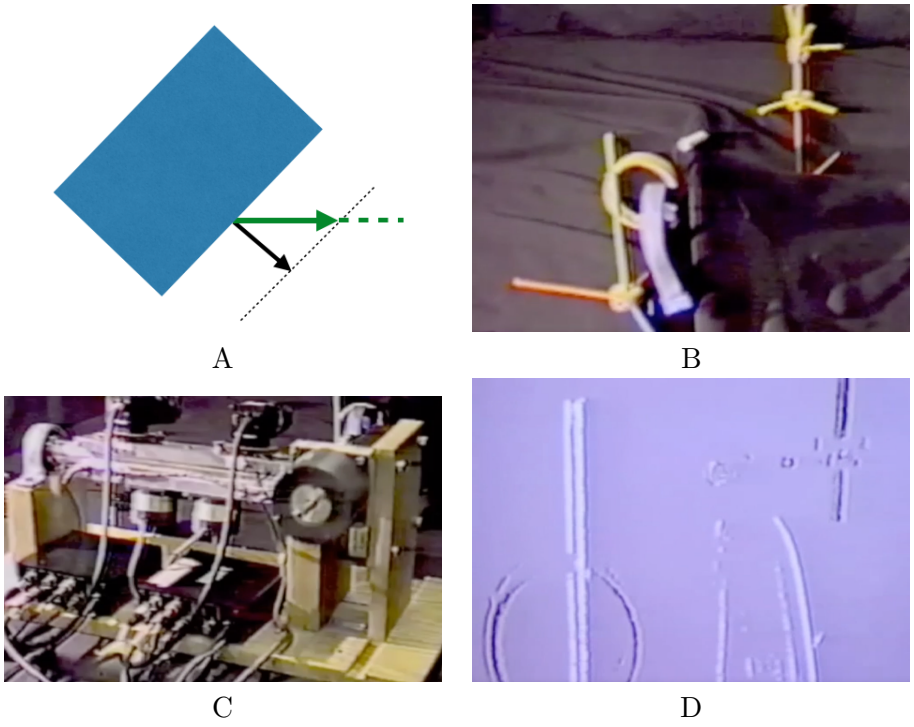

Figure 2: **Kinetic Depth Demonstration** A) Basic constraints. A basic horizontal motion of the head shown by a green arrow. The fundamental image constraint: motion perpendicular to the edge can be measured. Putting this projection together with an motor efference copy signal allows the true motion to be recovered B) Laboratory table scene with multicolored objects and a suspended piece of white chalk C) Robot head D) Kinetic signal recovered. During translated motion, the camera tracks the chalk with the result that points in front of the fixated chalk can be colored white and point behind can be colored black.

The active perception movement led to the construction of several laboratories' robot heads and the incorporation of the human repertoire of specific movements of saccades, vergence, and pursuit for basic scientific demonstrations as well as robot vision systems. However, the robot heads had

a much more significant impact than real-time dynamic image processing. Early vision theorizing had focused on a "bottom-up" perspective, wherein processing started at the retina and proceeded through brain regions. In contrast, the robots introduced "top-down" models with a focus moved to the agenda of the seeing agent.

The use of vision to solve such problems had been started earlier. This component was a focus for Just and Carpenter [1976] and Kosslyn [1980], but it was up to Ullman to make the explicit connection that this focus had been missing from the Marr paradigm [Ullman, 1987, Richards and Ullman, 1987]. Two of his examples appear in Fig.3. Both of these represent a problem that cannot be solved without calculation.

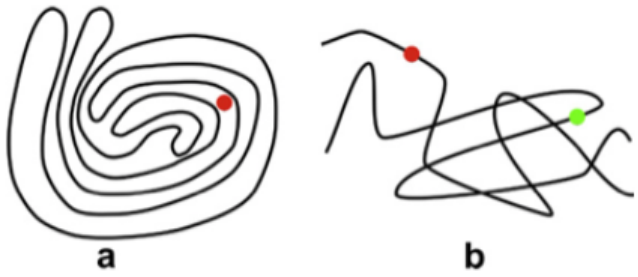

Figure 3: **Ullman's visual routines examples.** a) the question is to determine whether or not the red dot is inside or outside of the curve. This problem can be solved by counting the intersections on a line from the dot to the image edge. b) the question is whether or not the green dot and red dot are on the same curve. This problem can be solved by tracing the green dot along its curve and seeing if it meets the red dot. The essential is that these problems are outside of the realm of computing primal sketch properties.

While the problem-solving approach to vision was essential and timely, the advent of robot heads added a crucial component of platforms that could test constructive models. Subsequently, the visual routines model of human visual computation was been taken up in both the algorithms level and implementation level by several laboratories [Pahlavan et al., 1993, Aloimonos, 1990, Findlay et al., 2003]. In one paradigmatic example, Swain and Ballard [1991] showed that explicitly making a cortical parietal-temporal pathway distinction [Mishkin et al., 1983] led to significantly simpler algorithms for both object localization and identification.

Back to level I: Cortical hierarchies; the Bayes and Neural routines

Meanwhile, the discovery of regularities in the connection between cortical maps led to a complex of hierarchies [Van Essen et al., 1992] that revealed that adjacent maps had point-to-point connections. This revolutionary discovery was soon followed by mathematical formalisms to explain striate cortex receptive fields in terms of coding economies [Olshausen and Field, 1997], followed by an interpretation of the connections between these maps wherein cells' receptive field codes can be interpreted in terms of their ability to predict the cell responses of maps below in the hierarchy [Rao and Ballard, 1999]. Just as significant is the value of the error in prediction as a learning signal that can be used to adjust receptive field synapses. Another critical feature of this learning signal is that it can also be applied throughout the cortical hierarchy.

Given that these hierarchies have characteristic connectivity patterns, it opens up the possibility of another circuit interpretation, which is that the cortex is a vast compendium of behavioral experience. Thus we can invoke Bayes theorem to express a likelihood as:

$$P(W|S) = P(S|W)P(W)$$

with the following identifications:

1. W: the state of the world

2. S: the the sensory data

On the right-hand side, the first term is the likelihood, which records the probability of seeing the data given the state of the world, and the second term records the probability of a world state. Typically there is some constraint that guides the choice of probabilities for a world state that allows the likelihoods to be computed. A paradigmatic example comes from a motion illusion by Weiss et al. [2002] that has a ready Bayesian explanation is shown in Fig. 4.

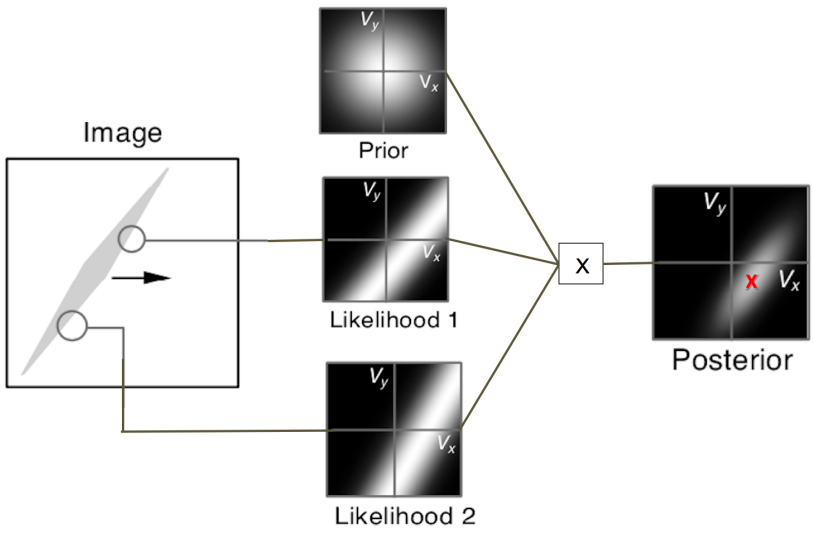

Figure 4: **Bayesian priors affect the perception of optic flow** (V_x, V_y)**.** A horizontally moving rhombus with $(V_0, 0)$ is perceived differently depending on its contrast. For a strong contrast, the motion would be seen as veridical, but for weak contrast shown, the perceived motion has a downward component indicated by the red X. This difference can be accounted for if the perceiving system has a preference for slow motion, indicated in the figure of by the distribution of optic flow vectors centered about zero motion. If local estimates of flow are sharp, they can overcome the prior bias, but if low contrast estimates are noisy, the maximum product of priors and likelihoods has a downward shift indicated by the red X in the posterior. (Redrawn from Weiss et al. [2002]).

A horizontally moving rhombus is rendered at different contrasts. In the space of optic flow values the constraints from the likelihoods of the linear edges appear as lines, but the low contrast case results in blurring. As a consequence, when combined with a prior for flows that prefers slow motions, the result of the Bayes prediction is for downward motion in the blurred case, congruent with the human perception.

The influence of the Bayes interpretations cannot be overestimated. In the first place, the perceptions do not have to be solely linked to sensory data but can be economically and flexibly expressed as the product of the likelihood and prior [Doya, 2007]. Secondly, Bayesian coding can be factored into all the levels in the cortical hierarchy. Without this perspective, the burden of the incoming image from the retina is to have an exacting code commensurate with the phenomenon of seeing; with it, the code just has to be sufficient to activate an appropriate network of priors.

Visual routines at the neural level Another development in the spirit was the demonstration of attentional effects at the neural level [Ito and Gilbert, 1999, Moran and Desimone, 1985, Ito and Gilbert, 1999, Tsotsos, 2011] had been the demonstration of more refined characterizations of visual routines, both in animals and humans. In a tracing task, Roelfsema showed evidence that a monkey solved a connectedness problem by mentally tracing a curve [Roelfsema et al., 2003]. By arranging to record from a single oriented cell on the path of a traced curve, he showed that the time course of the cell's elevated firing was consistent with the tracing's process.

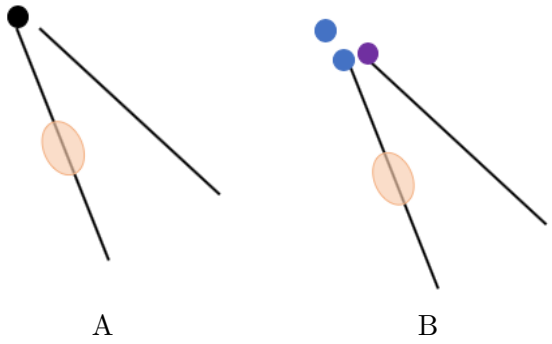

Figure 5: **Rolfsema's visual routines demonstration. A** A monkey is trained to hold fixation on a single black dot. Next, two lines appear. The monkey must make a saccade to the end of the line that contacts the fixation dots. In one of many trials, the experimenters are recording from a cell that has an appropriately oriented receptive field on the target line. The result is that spike activation is delayed for an interval consistent with mental tracing being used to determining the saccadic target. **B** In a more difficult task, the monkey, must pick the line that has the same color stub as the color of the fixation point. The hypothesis was that the time to resolve the correct target should take longer as two problems -color matching and tracing - have to be be solved. The result is a cell's activation is delayed appropriately.

III. The Awareness level: Agenda-based use of gaze behaviors

In introducing this level, we need to disambiguate it from level II, which can be similar in several ways. In the latter, the focus is on elucidating how visuomotor coordination is realized with no emphasis on cognition. Level II is exemplified by subconscious behaviors like one can tune out while driving a car on a familiar route and have no memory of the interval.

In contrast, Level III uses behaviors that could become subconscious, but a prime focus is on a subject's ability to plan, incorporate verbal instructions, and respond to on-off cues.

Head-free video-based eye tracking In primate vision, a central cue is the use of gaze fixation. Gaze control is impressively sophisticated, with six separate systems working in concert. All of these systems subserve the goal of stabilizing the gaze on the moving head, but of these, the use of saccades to fixate stationary targets and pursuit to track moving targets have a particular association with the subject's aims.

Early eye trackers were not head free and cumbersome to use. The dual Purkinje trackers required the subject to use a lab-mounted bite bar to stabilize the head during measurements [Van Rensbergen and De Troy, 1993]. Land's free head video system was much better but required the gaze to be recovered frame by frame using image post-processing, reminiscent of Yarbus's much earlier cornea-mounted mirror system [Yarbus, 1967]. Meanwhile, yet another technical advance made its appearance in the form of light-weight head-mounted eye-tracking. The importance of individual fixations had been long appreciated since Yarbus, but early head-restrained trackers significantly limiting experimental questions. Modern systems use pupil tracking via video to get a reliable eye in head signal to obtain an accuracy of $1 \sim 3$ visual degrees.

Yarbus's experiments showed that the gaze was intimately involved in the scene for the viewer's cognitive goals in scanning a scene and question answering. This new capability allowed this capability to deconstruct the progress of real-time visual analyses into their component tasks in hand-eye coordination tasks [Kowler et al., 1995, Gray and Boehm-Davis, 2000, Findlay et al., 2003]. One of the first of these showed that the coordination of a block copying task. Subjects shown a target pattern of a handful of colored blocks on a screen had to make a copy by selecting blocks from a reservoir with a cursor, drag them to a construction area and arrange them in an appropriate copy of the target. Fixations were used to direct the copying stages [Ballard et al., 1992]. Similar levels of coordination were observed in the sequential tapping of arrangements of

dots [Epelboim et al., 1995].

Land showed predictable visual coordination hand-eye coordination in driving [Land and Lee, 1994] and intercepting cricket pitches [Land and McLeod, 2000] as well as the more involved tasks in tea making, followed by Hayhoe's analysis of sandwich making [Land and Hayhoe, 2001]. Copying a pattern of blocks provided a moment-by-moment appreciation of the memory management of the process [Ballard et al., 1995].

Language and vision showed to synchronize Outside of the efforts in visual-motor coordination, the study of language had been limited to single word presentations and reading. Tanenhaus was quick to recognize the new capability's possibilities for situated language comprehension in the real world. In displays containing multiple objects, the actions asked of subjects could be studied in the context of the meaning of the objects and their phonemic content. His paper with Spivey linking fixations with question answering was ground-breaking [Tanenhaus et al., 1995]. In responding to directions, fixations were in lock-step with an utterance on a millisecond scale, a completely new finding. Situated behavior had been shown to have a comprehensive visual-audio embedding, but the tight temporal synchronization pointed to an underlying synchronization of neural computations across modalities.

This new paradigm had a huge impact and spawned many related research questions. In infant language learning, Chen and Smith showed that progress was intimately connected to cues from the care-giver given in real-world settings [Yu et al., 2005, Yu and Smith, 2012].

As for human studies during the early 2000s, the arrival of head-mounted displays incorporating eye trackers allowed human studies of gaze choices in virtual environments. Studies like that of [Droll and Hayhoe, 2007] showed that fixations could result in particular results. In a virtual reality block sorting task when the color of a held block was changed just before a sorting decision, subjects used the remembered color acquired during an earlier pickup rather than the changed color. This kind of result is very much in resonance with the Bayesian principles discussed earlier. Since the color had been inventoried at the beginning of the task using a visual routine, there was not a reason to check again as the prior probability is that objects typically don't change their color.

Mainstream reinforcement learning

Given that agenda-driven experience is ubiquitous, the next logical question is how such routines get programmed in the first place. The complete answer to this question is still open, but the general abstract answer is that routines are learned by reward [Kaelbling et al., 1996]. The neurobiological evidence for this in primates stems from Schultz [1998], which has been formally associated with reward by Schultz et al. [1997].

Reinforcement learning works with states, actions, transition functions and reward, as shown in a cursory example in Table The modern formalism of reinforcement algorithms was pioneered by

State	A state might be a phase in the preparation in a cooking recipe
Action	An action takes the chef from one phase to another
Transition function	A transition function describes the probability of transiting from a phase to another given a choice of action.

Table 2: Reinforcement learning learns a policy that specifies how to choose an action from any given state. Policies may be probabilistic. The main feature of the computation is that it computes actions based on expected discounted rewards.

Barto and Sutton, summarized in a classic text that has been recently updated [Sutton and Barto, 2018].

Thus progress in reinforcement learning has brought us to the point that dovetails with intellectual progress in cognitive science. The original questions of Marr in the 70s were confined to the early cortical areas, and neural models paved the way for the Bajczy's active vision that incorporated the extra degrees of freedom of the circumscribed vision system. The insights that the addition brought allowed a progression to even more formulations of embodied cognition, that integrated gaze and motor acts into cognitive models. At this point, this wave of progress has brought us to the point that we can start to revisit and update Newell's program of a "unified theory of cognition," and start to think what that model would look like, given all the legacy insights learned from vision studies.

Formal reinforcement learning seems like the outline of an answer to how the brain programs itself, but many details are unresolved. The biggest problem is that realistic reinforcement learning models scale exponentially with the size of the state space of the system. Handling large state space reinforcement models has been ameliorated with advances in deep learning [LeCun et al., 2015], and reinforcement learning has been sped up with improvements such as episodes [Botvinick et al., 2019], but given exponential scaling problems remain.

Without a way forward, there seems no way of adapting reinforcement as a model in cognition. Thus research is examining ways to decompose a more extensive problem into a set of modules specialized in solving more tractable smaller problems.

One view is that the next frontier needs to address how many streams of thought can be active at one time. This issue surfaced twenty years ago in neuroscience as the "binding problem," denoting the difficulty of parsing the interconnected neural substrate into usefully distinct components, but has lain fallow since Von der Malsburg [1999]. However, more recently, given a more cognitive context, interest in multiplexing has resurfaced [Panzeri et al., 2010]. Alternate motor plans are simultaneously active before choosing between them [Cisek, 2012]. This result resonates with another from the Tanenhaus paradigm, where it was shown that alternate object choices with names that had overlapping phonemes interfered on the millisecond level [Salverda et al., 2007]. One interpretation of these speedy responses is that the names were co-active at the neural level, as observed by Cisek. Even more remarkable, analyses of cortical recordings suggest that in a population of spikes, only 14% are used to represent task features, suggesting that the larger 86% are being used for other processes [Kobak et al., 2016]. In addition, there is a growing general interest in neuroscience-based multiplexing proposals [Akam and Kullmann, 2014].

Given the experimental evidence for multiple simultaneous reinforcing "threads, " researchers have focused on ways to compress the exponential space needed to account for behavior. One long-standing approach has been that of Sutton's *Options* [Sutton et al., 1999, Precup et al., 1998, Stolle and Precup, 2002], which associates the result of a program with its initial states of a lengthy coding of the internal state descriptions of how to bring about the result. Another is that of constraining the state space into parsimonious descriptions of small *Modules*. This tack defines modules a priori [Sprague et al., 2007], as depicted abstractly in Fig. 6.

It turns out that modules defined this way have several important properties that speak directly to cognitive concerns so that they are touched upon here.

Multiprocessing has a long-standing tradition of study in psychology in the form of dual tasks, but the influence of module reinforcement learning formalisms in multi-tasking is significant. The modules approach can serve as the underpinning for a "cognitive operating system" that can manage ongoing behaviors. The central assumption is that at any instant, the behavioral demands can be factored into a small set of active subsets. An example could be cooking the components of a meal. The different components on a stove require temporal attention as to their progress.

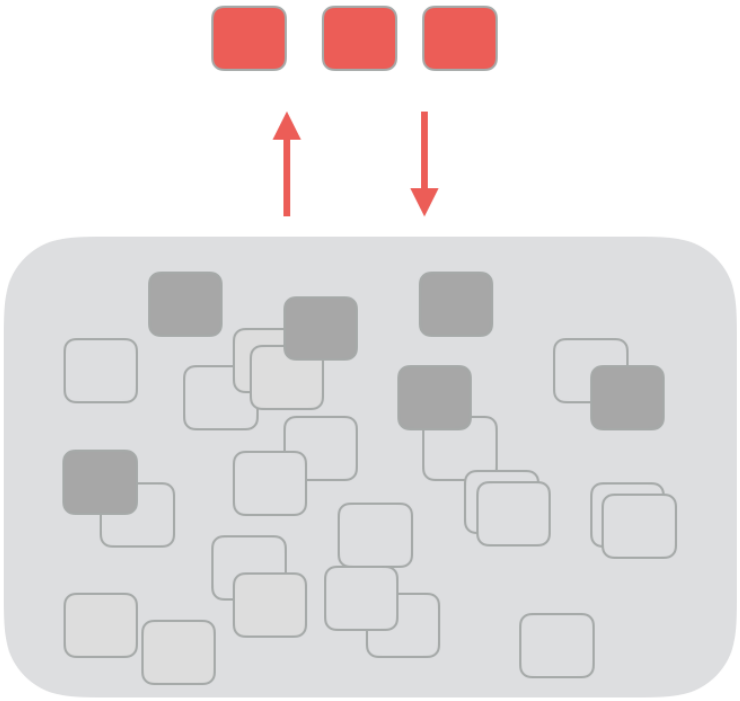

Figure 6: **Modules hypothesis** A general abstract module of behavior consists of enormous numbers of self-contained sensorimotor behaviors that are used in small numbers of compatible subsets to avoid neural cross-talk. Most often, these behaviors would be automatic, but if they have special activation constraints, they would require working memory. Thus working memory would be analogous to the computer science concept of "threads,' the state needed to keep track of an ongoing computational process. Modules can be activated and deactivated on a 300-millisecond timescale to track environmental reward opportunities. Shown in dark grey are modules that would be appropriate candidates for the current context.

While a small number of modules can be simultaneously active, the composition of the modules can be changed quickly, allowing for a rich tapestry of different modules to be active over time. The modal fixation interval of 200 to 300 milliseconds provides a logical time constant for module switching. The temporal agility of the operating system context allows complex behaviors to be addressed. In the example of driving, the car radio can be turned off in a situation of congested traffic.

A particular specialization for modules occurs when the action taken is shared by these modules. A typical case is accounting for the allocation of gaze for sets of modules that have spatially separated targets of interest. In this setting, multiple modules are co-active [Sprague et al., 2007]. Their state is drifting in time with an attendant increase in loss of reward as the policies cannot be as good in the uncertain state estimates. Given that two modules' states cannot be simultaneously updated, which one should get the gaze? A driving study [Johnson et al., 2014] showed that for each module, the expected loss in reward for not looking could be computed, and the module with the highest expected loss should have its state updated with a fixation. The model accounted for the distribution of human gaze intervals among multiple simultaneous behaviors, as shown in Fig. 7.

The dynamics of behavior management are the most significant advantage of the modules operating system approach, but here are two of several others.

One is that there is a ready solution for inverse reinforcement learning [Ng and Russell, 2000, Lopes et al., 2009, Neu and Szepesvári, 2007]. A significant component of learned experience comes from watching others. Given sequences of state-action pairs, the task of inverse reinforcement learning is to recover the reward distributions that were used to generate them. For many situations, one can assume that the watcher has similar basic modules as a demonstrator, but lacks the knowledge of how they are used in combination. The straightforward modules formulation of the problem assumes the demonstrator is using a weighted combination of modules and the problem for the learner is that of determining these weights from samples of the demonstrator's behavior. A Bayesian formulation results in exposing the weights in a way that they are easily recovered. In

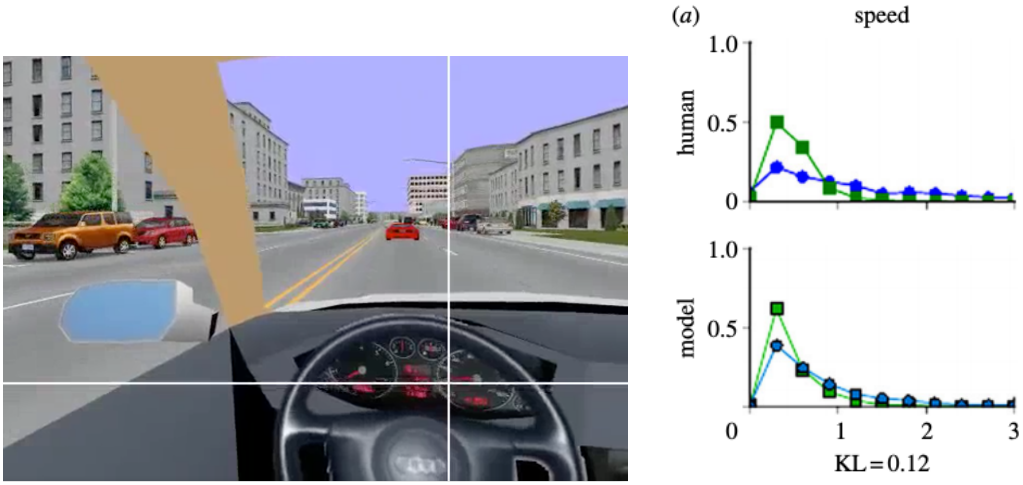

Figure 7: **A driving setting in virtual reality allows the study of multiplexing**. A subject tasked with following a red car and keeping the speed at 30 mph shares gaze between the speedometer and the followed car. At the instant shown gaze, indicated by the cross-hairs, is reading the location of the dial indicator for speed. Successfully executing both tasks can be achieved if separate modules for them are active and continually use gaze to update their state estimates of speed and car location.

a sidewalk navigation setting, Rothkopf and Ballard [2013], Zhang et al. [2018] were able to show that this works well for a navigation task replete with obstacles, but the method would generalize to any modularizable setting.

For the other example of the power of the modular setting, consider the issue the brain has in calculating the reward values of all its active sets of modules. Given that behaviors are selected on the basis of reward, how can the appropriate reward values be calculated? The modules formalization has an answer to this question if one assumes that active modules have access to the total reward earned at every instant, but are not given explicit information as to how to divide it up. In this setting at each instant a module r_i has two estimates of its reward. One is its running estimate, and another is the instantaneous reward handed out minus the estimates of the other modules in the currently active module cabal. Thus the estimate can be updated as the weighted sum of these two using the weight parameter β:

$$\hat{r}_i := (1-\beta)\hat{r}_i + \beta\Big(\sum_j r_j - \sum_{j\neq i}\hat{r}_j\Big)$$

This formula is very mathematically well-behaved and takes particular advantage of the expectation that over time modules will be active in different combinations [Rothkopf and Ballard, 2010]. The methodology is elegant and awaits substantive verification, but the importance of being able to bootstrap credit assignment online should not be underestimated.

In summary, the modules formalism is one approach to define cognitive control at the abstract level. It is not unique and has its modern-day competitors such as Sutton et al. [1999] and its temporally distant relative such as Newell and Allen [1994] and Anderson [1996] that are more programming language based. Nonetheless, these approaches share the goal of defining the execution of cognition at an abstract level.

The future

This review has described a progression towards understanding human vision. In retrospect, research has evolved a modeling perspective of hierarchies of algorithms to deal with brain issues at its different levels in the style of Newell's levels [Newell and Allen, 1994, Ballard, 2015]. However, for organizing the most open questions, Marr's neural level is still active and significant, given the several difficult questions that are very much open.

One such question is how to define the interface between automatic behaviors and modifiable behaviors. The beginning of visual models focused on a compartmental approach in characterizing the computations of early vision. The introduction of embedding of these formalisms in an active vision setting with the complete complement of multiple gaze modes as well as hand, head, and body contexts has had produced robust solutions to complete sensorimotor behaviors. These systems are becoming large-scale [Jain et al., 2017], but the focus of the embodied level is automatic behaviors. Such behaviors are valuable for their owners, owing to their economic overheads.

In contrast the awareness level adds effort that can be directed towards flexibly modifying the control of behaviors, with innovative abstract computational instantiations that can make full use of unexpected local world context, adding modifying state to code new behaviors.

Automatic behaviors are a mainstay, but human behavior uses both automatic and awareness levels. Thus this distinction is being given increased focus by many theorists, notably Graziano [2013], Dehaene [2014].

Into this rapidly evolving characterizations, Tanenhaus has made innovative demonstrations that language's rapid decision-making and production moves in a lock-step coupling with the perceptual-motor apparatus. This research program has made an enormous impact that is far from over, but this new venue brings up the relevance of modules [Sprague et al., 2007, Ballard et al., 2013, Zhang et al., 2018, Van Seijen et al., 2017]. Their promise can be illustrated with the example of multiplexing that brain computation modelers are starting to tackle. Silicon computers routinely take advantage of their high processing speeds to produce real-time multiprocessing by sharing small processing instants between multiple programs. It seems implausible that the brain would not have a solution as well, even if its space-time form was very different.

A multiplexing issue revolves around choosing good behaviors from a plethora of choices. That humans are addressing this problem can be appreciated from the study of "change blindness," the phenomenon that humans can be oblivious to significant changes to their environment [Simons and Levin, 1997]. But incompletely opposite to its moniker, change blindness is a hugely important filter that ignores vast numbers of unimportant or irrelevant distractors. This everyday human venue wherein a relatively small number of reward possibilities are chosen from the vast number of alternatives stands in contrast to some situations wherein *all* exigencies have to be handled.

Given the reality of a massive number of choices, a related important question is: How do we determine what events can interrupt our cognitive agenda? This question is unanswered, but an ingenious suggestion comes from a study by Roth et al. [2016]. A rat running on a treadmill is subjected to two different sets of visual stimuli. In one set, the stimuli are appropriate for the rat's motion. In another set, the visual motion is disjunct. It turns out the different cases are routed to the cortex using separate thalamic neural "buses." What the rat can predict is one case, and what that the rat cannot is another. Thus a hugely valuable insight is that things that can interrupt cognition may be those associated with being unpredictable and are handled by a particular anatomical pathway that registers things that cannot be predicted *with respect to the current cognitive agenda*. The study by [Triesch et al., 2003] also gives some insight here: task relevance of a changed feature influences if this change will be noticed.

A final and easy prediction to make is the continuing importance of the Deep Learning revo-

lution [Sejnowski, 2018, LeCun et al., 2015, Schmidhuber, 2015], with its evolving focus on human capabilities [Merel et al., 2019, Mnih et al., 2015]. Deep learning allows much more exacting real-time cognitive tests to be explored with brain-like architectures.

Although deep learning is proven to be extraordinarily powerful, many open questions are specifying an interface with the human neural system. The effort to a formative interface has been already started [DiCarlo et al., 2012, Mhaskar and Poggio, 2016] but still need to tackle the details of neural computation at millisecond levels [Vinck and Bosman, 2016, Maris et al., 2016, Fries, 2015, Panzeri et al., 2010] and rationalize the various forebrain oscillations. It has now been established that the forebrain uses meshed oscillations at least five frequencies - theta, alpha, beta, and gamma in the course of its computation [Vinck et al., 2010, Landau et al., 2015, Zheng and Colgin, 2015]. A coherent account of their interactions will be necessary for understanding at the neural level, as it ultimately will be necessary for understanding cognition.

REFERENCES

Thomas Akam and Dimitri M Kullmann. Oscillatory multiplexing of population codes for selective communication in the mammalian brain. *Nature Reviews Neuroscience*, 15(2):111, 2014.

John Aloimonos. Purposive and qualitative active vision. In *[1990] Proceedings. 10th International Conference on Pattern Recognition*, volume 1, pages 346–360. IEEE, 1990.

John R Anderson. Act: A simple theory of complex cognition. *American psychologist*, 51(4):355, 1996.

Ruzena Bajcsy. Active perception. *Proceedings of the IEEE*, 76(8):966–1005, 1988.

D. Ballard, M. Hayhoe, and J. Pelz. Memory representations in natural tasks. *J Cognitive Neuroscience*, pages 66–80, 1995.

Dana Ballard, Dmitry Kit, Constantin A. Rothkopf, and Brian Sullivan. A hierarchical modular architecture for embodied cognition. *Multisensory Research*, 26:177–204, 2013.

Dana H Ballard. *Brain computation as hierarchical abstraction*. MIT Press, 2015.

Dana H Ballard, Mary M Hayhoe, Feng Li, and Steven D Whitehead. Hand-eye coordination during sequential tasks. *Philosophical Transactions of the Royal Society of London. Series B: Biological Sciences*, 337(1281):331–339, 1992.

Dana H. Ballard, Mary M. Hayhoe, Polly K. Pook, and Rajesh P. N. Rao. Deictic codes for the embodiment of cognition. *Behavioral and Brain Sciences*, 20, 1997.

Harry Barrow, J Tenenbaum, A Hanson, and E Riseman. Recovering intrinsic scene characteristics. *Comput. Vis. Syst*, 2:3–26, 1978.

Matthew Botvinick, Sam Ritter, Jane X. Wang, Zeb Kurth-Nelson, Charles Blundell, and Demis Hassabis. Reinforcement learning, fast and slow. *Trends in Cognitive Sciences*, 23(5):408–422, 2019/08/02 2019. doi: 10.1016/j.tics.2019.02.006. URL `https://doi.org/10.1016/j.tics.2019.02.006`.

Christopher M Brown. The rochester robot. Technical report, ROCHESTER UNIV NY DEPT OF COMPUTER SCIENCE, 1988.

Paul Cisek. Making decisions through a distributed consensus. *Current opinion in neurobiology*, 22(6):927–936, 2012.

Andy Clark. *Supersizing the mind: Embodiment, action, and cognitive extension*. OUP USA, 2008.

Stanislas Dehaene. *Consciousness and the brain: Deciphering how the brain codes our thoughts*. Penguin, 2014.

James J DiCarlo, Davide Zoccolan, and Nicole C Rust. How does the brain solve visual object recognition? *Neuron*, 73(3):415–434, 2012.

Kenji Doya. *Bayesian brain: Probabilistic approaches to neural coding*. MIT press, 2007.

Jason A Droll and Mary M Hayhoe. Trade-offs between gaze and working memory use. *Journal of Experimental Psychology: Human Perception and Performance*, 33(6):1352, 2007.

Julie Epelboim, Robert M Steinman, Eileen Kowler, Mark Edwards, Zygmunt Pizlo, Casper J Erkelens, and Han Collewijn. The function of visual search and memory in sequential looking tasks. *Vision research*, 35(23-24):3401–3422, 1995.

John M Findlay, John M Findlay, Iain D Gilchrist, et al. *Active vision: The psychology of looking and seeing*. Oxford University Press, 2003.

Pascal Fries. Rhythms for cognition: communication through coherence. *Neuron*, 88(1):220–235, 2015.

Wayne D Gray and Deborah A Boehm-Davis. Milliseconds matter: An introduction to microstrategies and to their use in describing and predicting interactive behavior. *Journal of experimental psychology: applied*, 6(4):322, 2000.

Michael SA Graziano. *Consciousness and the social brain*. Oxford University Press, 2013.

Berthold KP Horn and Brian G Schunck. Determining optical flow. *Artificial intelligence*, 17(1-3): 185–203, 1981.

David H Hubel and Torsten N Wiesel. Receptive fields, binocular interaction and functional architecture in the cat's visual cortex. *The Journal of physiology*, 160(1):106–154, 1962.

Minami Ito and Charles D Gilbert. Attention modulates contextual influences in the primary visual cortex of alert monkeys. *Neuron*, 22(3):593–604, 1999.

Laurent Itti and Christof Koch. Computational modelling of visual attention. *Nature reviews neuroscience*, 2(3):194, 2001.

Suyog Dutt Jain, Bo Xiong, and Kristen Grauman. Fusionseg: Learning to combine motion and appearance for fully automatic segmentation of generic objects in videos. In *2017 IEEE conference on computer vision and pattern recognition (CVPR)*, pages 2117–2126. IEEE, 2017.

Leif Johnson, Brian Sullivan, Mary Hayhoe, and Dana Ballard. Predicting human visuomotor behaviour in a driving task. *Philosophical Transactions of the Royal Society of London B: Biological Sciences*, 369(1636):20130044, 2014.

Marcel Adam Just and Patricia A Carpenter. Eye fixations and cognitive processes. *Cognitive psychology*, 8(4):441–480, 1976.

Leslie Pack Kaelbling, Michael L Littman, and Andrew W Moore. Reinforcement learning: A survey. *Journal of artificial intelligence research*, 4:237–285, 1996.

Dmitry Kobak, Wieland Brendel, Christos Constantinidis, Claudia E Feierstein, Adam Kepecs, Zachary F Mainen, Ranulfo Romo, Xue-Lian Qi, Naoshige Uchida, and Christian K Machens. Demixed principal component analysis of neural population data. *Elife*, 5:e10989, 2016.

Christof Koch and Shimon Ullman. Shifts in selective visual attention: towards the underlying neural circuitry. In *Matters of intelligence*, pages 115–141. Springer, 1987.

Stephen Michael Kosslyn. *Image and mind*. Harvard University Press, 1980.

Eileen Kowler, Eric Anderson, Barbara Dosher, and Erik Blaser. The role of attention in the programming of saccades. *Vision research*, 35(13):1897–1916, 1995.

Michael F Land and Mary Hayhoe. In what ways do eye movements contribute to everyday activities? *Vision research*, 41(25-26):3559–3565, 2001.

Michael F Land and David N Lee. Where we look when we steer. *Nature*, 369(6483):742, 1994.

Michael F Land and Peter McLeod. From eye movements to actions: how batsmen hit the ball. *Nature neuroscience*, 3(12):1340, 2000.

Ayelet Nina Landau, Helene Marianne Schreyer, Stan Van Pelt, and Pascal Fries. Distributed attention is implemented through theta-rhythmic gamma modulation. *Current Biology*, 25(17): 2332–2337, 2015.

Yann LeCun, Yoshua Bengio, and Geoffrey Hinton. Deep learning. *nature*, 521(7553):436–444, 2015.

M. Lopes, F. Melo, and L. Montesano. Active learning for reward estimation in inverse reinforcement learning. *Machine Learning and Knowledge Discovery in Databases*, pages 31–46, 2009.

Eric Maris, Pascal Fries, and Freek van Ede. Diverse phase relations among neuronal rhythms and their potential function. *Trends in neurosciences*, 39(2):86–99, 2016.

David Marr. Vision: A computational approach, 1982.

David Marr and Tomaso Poggio. Cooperative computation of stereo disparity. *Science*, 194(4262): 283–287, 1976.

Josh Merel, Matthew Botvinick, and Greg Wayne. Hierarchical motor control in mammals and machines. *Nature Communications*, 10(1):1–12, 2019.

Hrushikesh N Mhaskar and Tomaso Poggio. Deep vs. shallow networks: An approximation theory perspective. *Analysis and Applications*, 14(06):829–848, 2016.

Mortimer Mishkin, Leslie G Ungerleider, and Kathleen A Macko. Object vision and spatial vision: two cortical pathways. *Trends in neurosciences*, 6:414–417, 1983.

Volodymyr Mnih, Koray Kavukcuoglu, David Silver, Andrei A Rusu, Joel Veness, Marc G Bellemare, Alex Graves, Martin Riedmiller, Andreas K Fidjeland, and Georg Ostrovski. Human-level control through deep reinforcement learning. *Nature*, 518(7540):529, 2015.

Jeffrey Moran and Robert Desimone. Selective attention gates visual processing in the extrastriate cortex. *Frontiers in cognitive neuroscience*, 229:342–345, 1985.

Gergely Neu and Csaba Szepesvári. Apprenticeship learning using inverse reinforcement learning and gradient methods. In Proceedings of the 23 Conference on Uncertainty in Artificial Intelligence, pages 295–302, 2007.

Newell and Allen. *Unified theories of cognition*. Harvard University Press, 1994.

Andrew Y. Ng and Stuart Russell. Algorithms for inverse reinforcement learning. In *in Proc. 17th International Conf. on Machine Learning*, pages 663–670. Morgan Kaufmann, 2000.

Alva Noë. *Out of our heads: Why you are not your brain, and other lessons from the biology of consciousness*. Macmillan, 2009.

Bruno A Olshausen and David J Field. Sparse coding with an overcomplete basis set: A strategy employed by v1? *Vision research*, 37(23):3311–3325, 1997.

Kourosh Pahlavan and Jan-Olof Eklundh. A head-eye systemanalysis and design. *CVGIP: Image Understanding*, 56(1):41–56, 1992.

Kourosh Pahlavan, Tomas Uhlin, and Jan-Olof Eklundh. Active vision as a methodology. *Active perception*, pages 19–46, 1993.

Stefano Panzeri, Nicolas Brunel, Nikos K Logothetis, and Christoph Kayser. Sensory neural codes using multiplexed temporal scales. *Trends in neurosciences*, 33(3):111–120, 2010.

Doina Precup, Richard S Sutton, and Satinder Singh. Theoretical results on reinforcement learning with temporally abstract options. In *European conference on machine learning*, pages 382–393. Springer, 1998.

Rajesh PN Rao and Dana H Ballard. Predictive coding in the visual cortex: a functional interpretation of some extra-classical receptive-field effects. *Nature neuroscience*, 2(1):79, 1999.

Whitman Richards and Shimon Ullman. *Image Understanding 1985-86*. Intellect Books, 1987.

Pieter R Roelfsema, Paul S Khayat, and Henk Spekreijse. Subtask sequencing in the primary visual cortex. *Proceedings of the National Academy of Sciences*, 100(9):5467–5472, 2003.

Morgane M Roth, Johannes C Dahmen, Dylan R Muir, Fabia Imhof, Francisco J Martini, and Sonja B Hofer. Thalamic nuclei convey diverse contextual information to layer 1 of visual cortex. *Nature neuroscience*, 19(2):299, 2016.

Constantin A Rothkopf and Dana Ballard. Credit assignment in multiple goal embodied visuomotor behavior. *frontiers in Psychology*, 1:173, 2010.

Constantin A Rothkopf and Dana H Ballard. Modular inverse reinforcement learning for visuomotor behavior. *Biological cybernetics*, 107(4):477–490, 2013.

Anne Pier Salverda, Delphine Dahan, Michael K Tanenhaus, Katherine Crosswhite, Mikhail Masharov, and Joyce McDonough. Effects of prosodically modulated sub-phonetic variation on lexical competition. *Cognition*, 105(2):466–476, 2007.

Jürgen Schmidhuber. Deep learning in neural networks: An overview. *Neural networks*, 61:85–117, 2015.

Wolfram Schultz. Predictive reward signal of dopamine neurons. *Journal of neurophysiology*, 80 (1):1–27, 1998.

Wolfram Schultz, Peter Dayan, and P Read Montague. A neural substrate of prediction and reward. *Science*, 275(5306):1593–1599, 1997.

Terrence J Sejnowski. *The deep learning revolution*. MIT Press, 2018.

Daniel J Simons and Daniel T Levin. Change blindness. *Trends in cognitive sciences*, 1(7):261–267, 1997.

Nathan Sprague, Dana Ballard, and Al Robinson. Modeling embodied visual behaviors. *ACM Transactions on Applied Perception (TAP)*, 4(2):11, 2007.

Martin Stolle and Doina Precup. Learning options in reinforcement learning. In *International Symposium on abstraction, reformulation, and approximation*, pages 212–223. Springer, 2002.

Richard S Sutton and Andrew G Barto. *Reinforcement learning: An introduction*. MIT press, 2018.

Richard S Sutton, Doina Precup, and Satinder Singh. Between mdps and semi-mdps: A framework for temporal abstraction in reinforcement learning. *Artificial intelligence*, 112(1-2):181–211, 1999.

Michael J Swain and Dana H Ballard. Color indexing. *International journal of computer vision*, 7 (1):11–32, 1991.

Michael K Tanenhaus, Michael J Spivey-Knowlton, Kathleen M Eberhard, and Julie C Sedivy. Integration of visual and linguistic information in spoken language comprehension. *Science*, 268 (5217):1632–1634, 1995.

Benjamin W Tatler, Mary M Hayhoe, Michael F Land, and Dana H Ballard. Eye guidance in natural vision: Reinterpreting salience. *Journal of vision*, 11(5):5–5, 2011.

Anne M Treisman and Garry Gelade. A feature-integration theory of attention. *Cognitive psychology*, 12(1):97–136, 1980.

Jochen Triesch, Dana H Ballard, Mary M Hayhoe, and Brian T Sullivan. What you see is what you need. *Journal of vision*, 3(1):9–9, 2003.

John K Tsotsos. *A computational perspective on visual attention*. MIT Press, 2011.

Shimon Ullman. Visual routines. In *Readings in computer vision*, pages 298–328. Elsevier, 1987.

David C Van Essen, Charles H Anderson, and Daniel J Felleman. Information processing in the primate visual system: an integrated systems perspective. *Science*, 255(5043):419–423, 1992.

J Van Rensbergen and A De Troy. A reference guide for the leuven dual-pc controlled purkinje eyetracking system psych, 1993.

Harm Van Seijen, Mehdi Fatemi, Joshua Romoff, Romain Laroche, Tavian Barnes, and Jeffrey Tsang. Hybrid reward architecture for reinforcement learning. In *Advances in Neural Information Processing Systems*, pages 5392–5402, 2017.

Martin Vinck and Conrado A Bosman. More gamma more predictions: gamma-synchronization as a key mechanism for efficient integration of classical receptive field inputs with surround predictions. *Frontiers in systems neuroscience*, 10, 2016.

Martin Vinck, Bruss Lima, Thilo Womelsdorf, Robert Oostenveld, Wolf Singer, Sergio Neuenschwander, and Pascal Fries. Gamma-phase shifting in awake monkey visual cortex. *Journal of Neuroscience*, 30(4):1250–1257, 2010.

Christoph Von der Malsburg. The what and why of binding: the modelers perspective. *Neuron*, 24 (1):95–104, 1999.

Yair Weiss, Eero P Simoncelli, and Edward H Adelson. Motion illusions as optimal percepts. *Nature neuroscience*, 5(6):598, 2002.

I. Yarbus. *Eye movements and vision*. Plenum Press, 1967.

Chen Yu and Linda B Smith. Embodied attention and word learning by toddlers. *Cognition*, 125 (2):244–262, 2012.

Chen Yu, Dana H Ballard, and Richard N Aslin. The role of embodied intention in early lexical acquisition. *Cognitive science*, 29(6):961–1005, 2005.

Ruohan Zhang, Shun Zhang, Matthew H Tong, Yuchen Cui, Constantin A Rothkopf, Dana H Ballard, and Mary M Hayhoe. Modeling sensory-motor decisions in natural behavior. *PLoS computational biology*, 14(10):e1006518, 2018.

Chenguang Zheng and Laura Lee Colgin. Beta and gamma rhythms go with the flow. *Neuron*, 85 (2):236–237, 2015.