Learning to Attend — From Bottom-Up to Top-Down

Hector Jasso[1] and Jochen Triesch[2,3]

[1] Dept. of Computer Science and Engineering, University of California, San Diego
La Jolla, CA 92093, USA
[2] Frankfurt Institute for Advanced Studies, J.W. Goethe University
Frankfurt am Main, Germany
[3] Dept. of Cognitive Science, University of California, San Diego
La Jolla, CA 92093, USA

Abstract. The control of overt visual attention relies on an interplay of bottom-up and top-down mechanisms. Purely bottom-up models may provide a reasonable account of the looking behaviors of young infants, but they cannot accurately account for attention orienting of adults in many natural behaviors. But how do humans learn to incorporate top-down mechanisms into their control of attention? The phenomenon of gaze following, i.e. the ability to infer where someone else is looking and to orient to the same location, offers an interesting window into this question. We review findings on the emergence of gaze following in human infants and present a computational model of the underlying learning processes. The model exhibits a gradual incorporation of top-down cues in the infant's attention control. It explains this process in terms of generic reinforcement learning mechanisms. We conclude that reinforcement learning may be a major driving force behind the incorporation of top-down cues into the control of visual attention.

1 Introduction

When we look at a visual scene, we can only see a very small part of it at a high resolution at any time. This is because the retina, which converts the incoming light into electrical signals that are passed on to the brain, has only a very small region, the fovea, that samples the visual scene at high resolution. The fovea only represents the central two degrees of the visual field, corresponding to roughly twice the width of your thumb at arm's length, and the resolution falls off quickly outside of these central two degrees. Because our vision is accurate in only this small central region, we constantly have to move our eyes to aim the fovea at relevant visual targets for detailed analysis. On average, we engage in roughly 3 such eye movements per second, separated by fixation periods where the eyes remain more or less stable on one location. The brain processes that control what parts of a visual scene we overtly attend to in this way are extremely complex and only poorly understood. While it is possible to voluntarily control the movements of our eyes, most eye movements happen without our awareness of them.

1.1 Bottom-up vs. Top-down Control of Attention

A popular attempt to categorize the different mechanisms that control our overt visual attention is to distinguish bottom-up and top-down mechanisms. While bottom-up mechanisms are frequently characterized as automatic, reflexive, and fast, requiring only a comparatively simple analysis of the visual scene, top-down mechanisms are thought of as more voluntary and slow, requiring more complex inferences or the use of memory.

Bottom-up mechanisms are closely related to the idea of a saliency map [1]. A saliency map is a topographic map of the visual scene that encodes the visual conspicuity of different locations. Importantly, the conspicuity of a stimulus critically depends on the context in which the stimulus is embedded. For example, a red berry among green foliage would be highly salient. Often a saliency map is computed as the sum of contributions of a number of simple feature maps for, say, movement, color, and contrast. Attention is directed to the most salient location in this map, with an inhibition of return mechanism preventing the currently attended location from being attended again.

Top-down attention mechanisms have a much more diverse nature and their distinction from bottom-up attention mechanisms may not always be very clear. As a working definition, we require top-down attention to be based on significant analysis of the visual scene beyond the calculation of a visual saliency map based on elementary feature channels. Such elaborate visual analysis is performed by higher visual cortical areas and may draw on long-term and working memory processes. Since young infants have only very limited capacities for such elaborate analysis, their attention control is likely to be dominated by bottom-up mechanisms.

Usually, the deployment of visual attention via top-down mechanisms is strongly influenced by the demands of the current behavior. The classic evidence for this was provided by Yarbus, who showed that subjects looking at a picture will engage in very different fixation patterns depending on whether they are instructed to, say, estimate the ages of the people in the image, or to memorize all objects in the scene [2]. The importance of behavioral goals for the deployment of visual attention is even more obvious in tasks that require physical interaction with the enviroment, such as the manipulation of various objects during the preparation of a cup of tea [3, 4] or a sandwich [5].

While a comprehensive scheme for categorizing top-down attention mechanisms is beyond the scope of this chapter, at least the following mechanisms can be distinguished.

- Visual search. A stored represenation of the appearance of an object or object class (my keys, a horse, a red object) is used to direct attention to locations likely to contain a desired target. Sometimes this is thought of in terms of higher cortical areas modulating the gain of different feature channels during the computation of an otherwise bottom-up saliency mechanism [6].
- Motor Control. The eyes are moved to where they are needed to allow efficient motor control. For example, when driving a car around a curve, the driver will usually tend to look at the tangent point of the curve [3, 4].

- Prediction. The eyes are frequently moved to locations where an interesting event is predicted to happen. For example, six-month-old infants can already learn to predict where an interesting visual stimulus will occur and move their eyes to this location before the stimulus actually appears [7]. Similarly, cricket batsmen will predict the trajectory of the ball when it leaves the pitcher's hand and fixate the expected bounce point of the ball [3, 4]. Interestingly, expert batsmen will fixate this point earlier than novices.
- Memory of location. Frequently attention is directed towards objects that were looked at previously but only now have become relevant for the current task. For example, we may recall that we put down our cup on the table behind us and turn around to pick it up.
- Social Environment. Last but not least, in humans (and some other social species) the control of visual attention is in part driven by where other people are looking. For example, when your conversation partner suddenly turns to the side to stare at something, you are very likely to turn in the same direction to identify what he or she is looking at. This behavior is called *gaze following*, and its development in infancy allows a glimpse at how a specific top-down mechanism is acquired and incorporated into the control of visual attention. This will be the topic of the remainder of this chapter.

1.2 Gaze Following

The development of gaze following during infancy has been studied for over 30 years. The motivation behind much of this research is the desire to better understand in how far infants at various ages conceive other people as perceiving intentional agents, i.e. in how far they have developed of a "theory of mind". In typical gaze following experiments [10], an experimenter and an infant sit facing each other (see Fig. 1). A target object is placed on one side of the infant's midline, and a distracter object on the other side. The experimenter first catches the infant's attention and then turns towards the target, waiting in this position for a few seconds. It is noted if within this time the infant turns towards the target (considered a correct response), towards the distracter (considered an incorrect response), or does not turn at all (considered a non-response). Trials are repeated a number of times to estimate whether the infant is more likely turn to the object indicated by the turning of the caregiver's head and eyes than to the distracter.

Many variations of this experimental setup have been used to study the development of gaze following in infants. For example, it has been found that younger infants will not follow gaze if the objects are positioned behind them (so that they are outside their field of view), but older infants will [11–13]. Younger infants will also sometimes erroneously follow gaze to (extra) distracters positioned on the same side of the room as the target but not being looked at by the caregiver [11, 12]. A variation of this experimental setup involving conflicting head and eye direction cues was used to show that younger infants, but not older ones, will tend to disregard eye direction cues [14–16]. In [17], it was shown that autistic children exhibit diminished gaze following behaviors.

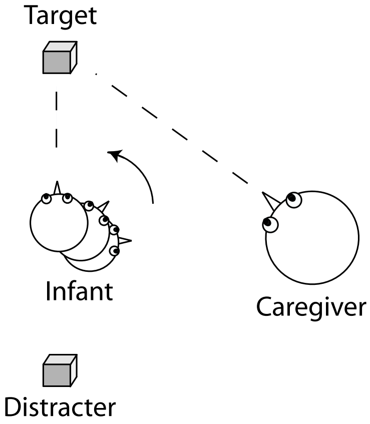

Fig. 1. Basic gaze following experimental setup.

Gaze following experiments are not usually discussed in the context of top-down and bottom-up attention integration, as is the case of typical visual search experiments [18]. Gaze following, however, can be seen as an instance of a visual search task: While in typical visual search experiments the subject is asked to locate an object among distracters within a visual scene, in gaze following the object can be outside the subject's field of view, and therefore outside the initial visual scene (e.g. [11–13]). Another difference is that the goal is not explicitly given, but instead is implicit in the experimenter's gaze direction.

In the last few years, robotic [19–24] as well as purely computational models [25–27] of gaze following have been proposed. However, they have not been analyzed from the perspective of bottom-up and top-down attention. In addition, while they are discussed in the context of infant development, they have not been used to replicate the experiments described above[4], making it impossible to compare them to empirical evidence.

In the following we present a recent computational model of the development of gaze following in infancy [47] that has been used to successfully replicate many of the experimental findings about the development of gaze following in infants. Of particular interest for the current discussion is how the model learns to combine bottom-up and top-down mechanisms to locate rewarding visual targets.

[4] Although [23] claims to replicate the developmental stages described in [11, 12], there seem to be some inconsistencies in their mapping of developmental stages between the model and the experimental observations.

2 Methods

Our model of the development of gaze following is based on reinforcement learning [8, 9]. The basic idea is that infants prefer to look at salient or otherwise interesting visual stimuli because these will trigger a reward signal in the infant's brain. If the assumption is made that other people (in partcular the infant's caregivers) also tend to look at rewarding visual stimuli, then the infant can learn where to find such stimuli by observing where the others are looking. The infant learns that a caregiver looking in a certain direction is often associated with an interesting object or event occurring somewhere along the caregiver's line of sight. The model has been used to replicate major aspects of the developmental trajectory of gaze following as described in the previous section. A comprehensive description of the model is given in [47]. Here we focus on the aspect of bottom-up versus top-down control of attention.

2.1 Model description

Modeling the environment The environment is modeled as follows: Infant and caregiver are positioned facing each other with a 40 cm separation between them in a two-dimensional environment as illustrated in Fig. 1. Objects can be placed anywhere except at the same location as the infant or caregiver. Time is discretized into steps of 1 second.

Infant visual system The infant's visual system comprises three different components (see Fig. 2 left): a saliency map ($\mathbf{s}$), a head direction detector ($\mathbf{h}$), and an eyes direction detector ($\mathbf{e}$).

Saliency Map ($\mathbf{s} = [s_1, ..., s_{96}]$) Indicates the presence of visual saliency in a body-centered coordinate system with 96 different regions in space, along 24 heading ranges and 4 depth ranges. Heading 1 corresponds to heading angles between -7.5° and 7.5°, heading 2 corresponds to angles between 7.5° and 22.5°, and so on, covering all 24 different headings. Depth 1 corresponds to distances (from the infant's perspective) of up to 0.8 meters away, depth 2 corresponds to distances of 0.8 to 1.2 meters, depth 3 corresponds to distances of 1.2 to 1.7 meters, and depth 4 corresponds to distances of more than 1.7 meters.

The saliencies of objects and caregiver within the infant's field of view are added to the element in $\mathbf{s}$ corresponding to their location (heading and depth), after foveation and habituation are calculated: Foveation causes an object's perceived saliency to decay as it falls outside the infant's center of vision according to the following formula (adapted from the contrast sensitivity function proposed by [28]):

$$foveation(\theta) = 0.2 + 0.8 \frac{1}{1 + k_{Ecc} \cdot \theta} , \tag{1}$$

where θ is the eccentricity in visual angle of the object, and k_{Ecc} is a constant that defines how the sensitivity diminishes with eccentricity. The offset of 0.2

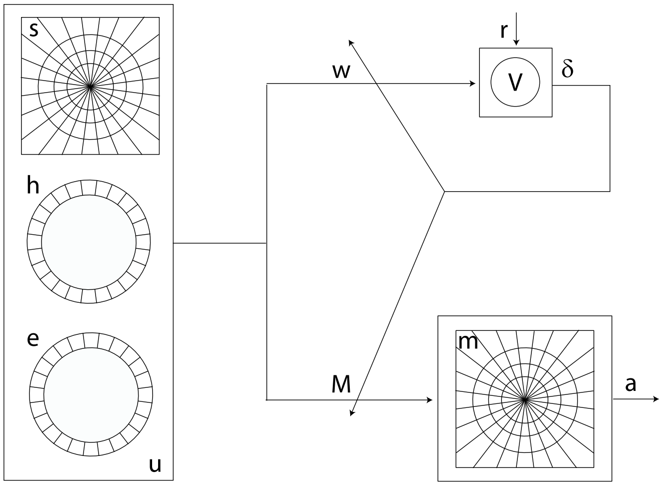

Fig. 2. Model architecture. Details of the infant visual system (left) and of the actor-critic reinforcement learning model (right). Features calculated from the *Saliency Map* **s**, *Caregiver Head Direction* **h**, and *Caregiver Eyes Direction* **e** are combined into **u**, weighted using **w** and added into V to estimate the value of the present state. They are also weighted using **M**, added into **m**, and passed through a softmax selection formula to choose the next action a.

prevents values from decaying to close to zero when objects are in peripheral vision (i.e. "in the corner of the eye"), which helps replicate some of the gaze following experimental results where a distracter object in the periphery of vision captures the attention of the infant.

The infant habituates separately to each object, according to the discretized version of the following exponential decay formula proposed in [29]:

$$\tau_H \frac{d\phi_{o_j}(t)}{dt} = \alpha_H(\Phi_{o_j} - \phi_{o_j}(t)) - S_{o_j}(t) \ , \tag{2}$$

where $\phi_{o_j}(t)$ is object j's habituated saliency at time t and Φ_{o_j} its original, dishabituated, saliency; $S_{o_j}(t)$ is equal to Φ_{o_j} if the infant is looking at object j at time t and 0 otherwise; τ_H is a time constant that specifies the rate of habituation (a smaller τ_H resulting in faster habituation); and α_H controls the level of long-term habituation. A similar formula applies for ϕ_C and ϕ_I, the habituated saliencies of the caregiver (Φ_C) and the infant (Φ_I), respectively.

The caregiver's saliency is halved when the caregiver's face is in profile (i.e. looking away from the infant). This reflects infant's preference for looking at other's gaze directed at them than diverted elsewhere [30].

Finally, when an element s_i of **s** is outside the infant's field of view, its new value is calculated by multiplying the previous value by a constant d ($0 < d < 1$), a "memory decay" factor. This enables the model to temporarily remember recently observed states of the world.

The exact formula for calculating **s** is: $s_i = S_O + S_C + S_{M_i}$, where

- $S_O = \sum_{j=1}^{N} S_{o_j}$ and $S_{o_j} = \phi_{o_j} foveation(\theta_{o_j})$ if o_j is within the infant's field of view, 0 otherwise, θ_{o_j} being the angular distance of the object from the center of vision,
- $S_C = \phi_C foveation(\varphi_I)$ if the caregiver is present and withint the infant's field of view; 0 otherwise,
- $S_{M_i} = s_i(t-1)d$ if the location is outside the infant's field of view, 0 otherwise.

The primary visual cortex has been proposed as an instantiation of a saliency map [31, 32]. Our assumption of a body-centered representation (in contrast to a retinotopic one) is not physiologically accurate but it frees us from having to model coordinate transformations between different coordinate systems (although it is an interesting question in its own right when and how infants learn to compute certain coordinate transformations).

Head Direction Detector ($\mathbf{h} = [h_1, ..., h_{24}]$) Indicates 24 possible caregiver head directions as perceived by the infant. Heading ranges are similar to those in **s**. If the infant is looking at the caregiver, the value of each h_i is calculated according to an exponential decay, so that the closer h_i is to the caregiver's heading, the higher the value. **h** is normalized (using linear scaling) so that the sum of all h_i add to 1. This decay is gentler at the beginning of learning, and gets progressively sharper with time, to reflect the development of this ability from infancy [30, 10–12, 33] to adulthood [34–36].

If the infant is not looking at the caregiver, then the values of **h** are calculated by multiplying the previous value by the memory constant d, (the same as in the calculation of **s**), to enable the model to temporarily remember recently observed head directions of the caregiver.

The exact formula for calculating **h** is: $h_i = H_C + H_{M_i}$ with a posterior scaling of all h_i so that $\sum_{i=1}^{24} h_i = 1$, where

- $H_C = exp(-(\varphi_H - \theta_{I_i})^2/\sigma_H{}^2)$ if the caregiver is present and the infant is looking at the caregiver, φ_H being the caregiver's heading direction, θ_{I_i} the angle corresponding to heading i's center ($\theta_{I_1} = 0°$, $\theta_{I_2} = 15°$, $\theta_{I_3} = 30°$, ... $\theta_{I_{24}} = 345°$), and σ_H a parameter that specifies the exponential decay; 0 otherwise;
- $H_{M_i} = h_i(t-1)d$ if the caregiver is absent or outside the infant's field of view, 0 otherwise.

Eyes Direction Detector ($\mathbf{e} = [e_1, ..., e_{24}]$) Similar to **h**, but computed with the caregiver's eye direction instead of head direction, and with a different exponential decay parameter (σ_E instead of σ_H). Additionally, when the caregiver is present and within the infant's field of view but turning back, all values e_i are set to zero. This reflects the fact that when the caregiver is facing backwards with respect to the infant, the eyes are not visible.

Such representations of head and eye direction may be found in the superior temporal sulcus (STS) in monkeys, and are likely to exist in humans, too [37].

Separate mechanisms for the caregiver's head pose and eye direction allow us to capture the development of the infant's differential sensitivity to these cues.

Reinforcement Learning Model The infant's visual system serves as input to an actor-critic reinforcement learning system [8] that drives actions. The *critic* (see Fig. 2, upper right) approximates the value of the current state as $V(t) = \mathbf{w}(t)\mathbf{u}(t)$ where $\mathbf{w}(t) = (w_1(t), w_2(t), ..., w_{N_s}(t))$ is a weight vector, $\mathbf{u}(t) = (\mathbf{s}(t), \mathbf{h}(t), \mathbf{e}(t))^T$ is the value of the input features from the visual system at time t, and N_s is the number of features ($N_s = dim\ \mathbf{s} + dim\ \mathbf{h} + dim\ \mathbf{e} = 96+24+24 = 144$). The weight vector $\mathbf{w}(t)$ is updated according to the formula:

$$\mathbf{w}(t+1) = \mathbf{w}(t) + \eta\delta(t)\mathbf{u}(t) \ , \tag{3}$$

where η is the learning rate, and $\delta(t)$ specifies the temporal difference error, defined as the difference between the immediate reward received plus the estimated future discounted reward, minus the current estimated value of the state:

$$\delta(t) = r(t) + \gamma V(t+1) - V(t) \ , \tag{4}$$

where $r(t)$ is the reward at time t, $V(t+1)$ the estimated value of the new state after taking the action, and γ the reward discount factor.

The *actor* (see Fig. 2, lower right) specifies the action to be taken, directing the infant's attention to one of 24 possible different headings and one of four different depths, with a total of $N_a = 96$ different possible actions ($A = (H, D)$, $H \in \{0°, 15°, 30°, ..., 345°\}$, $D \in \{0.4, 1.0, 1.45, 2.0\}$), where A is the action, and H and D are the heading and depth, respectively, where attention is directed to.) The action is chosen probabilistically according to the softmax decision rule:

$$P[a] = \frac{\exp(\beta m_a)}{\sum_{a'=1}^{N_a} \exp(\beta m_{a')}} \ , \tag{5}$$

m_a being the action value parameter for action a for the present state: $\mathbf{m} = \mathbf{M}\mathbf{u}$, where $\mathbf{M}$ has as many columns as there are input features and as many rows as there are actions. A higher value of m_a increases the chances of selecting action a. β is an "inverse temperature" parameter which increases exploitation versus exploration with a larger value. $\mathbf{M}$ is updated according to:

$$M_{a'b}(t+1) \leftarrow M_{a'b}(t) + \epsilon(\delta_{aa'} - P[a'; \mathbf{u}(t)])\delta(t)u_b(t) \ , \tag{6}$$

where η is the same learning rate as above, $\delta(t)$ is the critic's temporal difference error (defined above), a is the action taken, $P[a'; \mathbf{u}(t)]$ is the probability of taking action a' at state $\mathbf{u}$(t), and $\delta_{aa'}$ is the Kronecker delta, defined as 1 if $a = a'$, 0 otherwise.

Reward is obtained as the saliency of the position where attention is directed to after the action is taken and $\mathbf{s}$ updated with the result of the action (the value of $\mathbf{s}$ corresponding to the depth/heading of the selected a, but in the next time step, and with a foveation corresponding to the new infant heading). The definition of salience as reward is based on studies of infant visual expectations and the organization of their behavior around these expectations [7].

The firing of dopaminergic neurons in the ventral tegmental area has been associated with the temporal difference signal [38]. Layer **m** corresponds to a representation of pre-motor neurons, which are activated when an action is planned. Interestingly, layer **m** also shares some characteristics with so-called mirror neurons, which have been hypothesized to be implicated in imitation and action understanding. This topic is discussed in [39].

2.2 Training and testing scheme

The simulation starts with the infant and caregiver in the middle of the room facing each other. N_o objects are placed in the room, where N_o is drawn from a geometric probability distribution with average $\bar{N}_o$.

Objects are placed randomly around the infant with distances from the infant taken from a radially symmetric normal probability function with standard deviation of σ_o. The saliency Φ_{o_i} for each object i is drawn from an exponential probability with average $\bar{\Phi}_o$.

After a number of time steps drawn from a geometric probability function with average $\bar{T}_{objects}$, all objects in the room are removed and replaced by new objects, with positions and saliencies drawn randomly as described above. Additionally, after a number of time steps drawn from a geometric probability function with average $\bar{T}_{present}$, the caregiver leaves the room. The caregiver returns to continue interacting with the infant after a number of time steps drawn from another geometric probability function with average $\bar{T}_{absent}$.

The simulation is run for *10,000,000* training steps (roughly corresponding to 115 days of a wake infant), during which gaze following develops. During training, the caregiver always looks at the most salient point in the room, which in some cases will be the infant. The caregiver's perceived saliencies are mediated by the same foveation and habituation mechanisms (with identical parameters) as in the infant's visual system. The caregiver's head direction is slightly offset from that of the eyes according to a Gaussian distribution with $\sigma = 5°$ and $\mu = 0°$. This offset is recalculated for every gaze shift that the caregiver does. This reflects the fact that eyes and head are not always perfectly aligned, and corresponds to values observed in naturalistic settings [5]. The infant acts according to the reinforcement learning algorithm described above.

2.3 Parameter Setting

This section describes the default parameter values of the model. Table 1 summarizes the parameters and their settings.

Environment Modeling parameters: These parameters were set to simulate a naturalistic environment where caregiver and infant interact with each other in a fairly dynamic environment. This is based on assumptions about a structured environment as described in [40].

The infant's saliency as well as the caregiver's saliency is set to 4.0, while the average object saliency is set to 1.0. This makes the infant and caregiver above-average objects of interest. The caregiver's saliency is given a high value because newborns preferentially orient towards faces [41, 42], and because caregivers provide social contingency, which is preferred by infants [43]. With this parameter setting, most of the objects will be less salient than the caregiver or infant, with the possibility of having objects that are more interesting.

The average number of objects, $\bar{N}_o$, is set to 4, for a reasonably rich environment (this value can be set lower (but not to 0) or higher without significant differences in results). σ_o, the object placement spread, is set to 1.0 m. This samples all the four depths in the infant's visual system with roughly the same frequency. These values are intended to simulate a setting such as a nursery where objects are placed around the infant for it to play with, but with some objects like walls, doors, desks, or chairs far away.

$\bar{T}_{present}$ is set to 60 seconds, $\bar{T}_{absent}$ to 60 seconds, and $\bar{T}_{objects}$ is set to 5 seconds. This models a fairly dynamic environment, with typical object displacements such as the caregiver manipulating a toy in front of the infant while playing or teaching, or the infant itself manipulating the objects. Having the caregiver present half the time simulates the substantial time involved in child rearing, which includes activities with face-to-face interaction between infant and caregiver, such as feeding and playing.

Infant visual system parameters: FOV, the infant's field of view, is set to 180°, simulating the human visual system. Habituation's τ_H and α_H are set to 2.0 and 1.0 respectively, resulting in almost complete habituation after about 5 seconds.

The initial value of σ_H ($\sigma_{H_{initial}}$) is set to 50°. σ_H decrements 5° ($\sigma_{H_{step}}$) every 200,000 time steps, reaching a final value ($\sigma_{H_{final}}$) of 1°. The corresponding values for $\sigma_{E_{initial}}$, $\sigma_{E_{step}}$, and $\sigma_{E_{final}}$ are 50°, 2°, and 1°. This corresponds to an eye direction signal more difficult to interpret than the head direction cue (the eyes being smaller than the head), and allows us to replicate experiments where the value of the other's eye direction is learned slower than that of the head direction. These settings are important to replicate a gradual incorporation of the eyes direction cues.

Reinforcement Learning parameters: In general, these parameters are set so that learning can take place fast, but not so fast that learning becomes unstable.

The learning rate η is set to 0.01 for smooth learning. The discount factor γ is set to 0.1. The "inverse temperature" parameter β is set to 30, resulting in a high level of exploration early on, and a fairly "greedy" action selection afterwards, as the weight values of $\mathbf{w}$ and $\mathbf{M}$ increase through learning. All elements of $\mathbf{M}$ and $\mathbf{w}$ are initialized to zero, reflecting an absence of previous experience with saliencies and gaze, and thus of any innate gaze following abilities.

On using a single set of parameters: The model exhibits two characteristics that make it appealing: First, a single parameter specification is sufficient to replicate a wide variety of gaze following experiments, as described below.

Table 1. Overview of model parameters, their allowed ranges and default values.

Symbol	Explanation	Range	Default
Environment modeling			
Φ_I	Infant's saliency	$(-\infty, \infty)$	4.0
Φ_C	Caregiver's saliency when facing the infant	$(-\infty, \infty)$	4.0
$\bar{\Phi}_O$	Average object saliency	$(-\infty, \infty)$	1.0
$\bar{N}_o$	Average number of objects	$[0, \infty)$	4
σ_o	Object placement spread around infant	$[0, \infty)$	1.0 m
$\bar{T}_{present}$	Average caregiver interaction interval	$[0, \infty)$	60 s
$\bar{T}_{absent}$	Average caregiver absence interval	$[0, \infty)$	60 s
$\bar{T}_{objects}$	Average object replacement interval	$[0, \infty)$	5 s
Infant visual system			
FOV	Size of field of view	$[0°, 360°]$	180°
σ_H	Head direction perception fuzzyiness	$(0°, \infty)$	
$\sigma_{H_{initial}}$	Initial σ_H value	$(0°, \infty)$	50°
$\sigma_{H_{final}}$	Final σ_H value	$(0°, \infty)$	1°
$\sigma_{H_{step}}$	Decrement in σ_H per 200,000 time steps	$[0°, \infty)$	5°
σ_E	Eyes direction perception fuzzyiness	$(0°, \infty)$	
$\sigma_{E_{initial}}$	Initial σ_E value	$(0°, \infty)$	50°
$\sigma_{E_{final}}$	Final σ_E value	$(0°, \infty)$	1°
$\sigma_{E_{step}}$	Decrement in σ_E per 200,000 time steps	$[0°, \infty)$	2°
τ_H	Habituation rate	$[0, \infty)$	2.5
α_H	Target of habituation	$[1.0, \infty)$	1.0
d	Memory decay factor	[0,1]	0.5
Reinforcement Learning			
η	Learning rate	$[0, \infty)$	0.01
γ	Discount factor	$[0, \infty)$	0.1
β	Inverse temperature	$[0, \infty)$	30

This leads to a stronger claim of fitting the data than the alternative method of using different parameter settings for different experiments of the same phenomena [44]. Second, the model can still replicate the experiment results even with reasonable modifications to these parameters. For example, having the caregiver present less often slows down learning, but does not preclude it. The same can be said about the complexity of the environment in the model.

It should be noted that many of these parameters were introduced because of a desire to replicate as many experiments as possible such as spatial characteristics of the room, the different head and eye directions, a limited field of view, etc. And while simpler versions of the model could be used to drive the point for different experiments (for example, the limited field of view is not necessary to replicate experiments investigating the different effect of eye and head direction), there is value in having a single model with a single set of parameters.

3 Results

3.1 Replicating the development of gaze following in infancy

The model replicates the major aspects of the developmental trajectory of gaze following in infants [47]. For example, at first it does not follow gaze to objects outside its field of view, but does so at later stages [45]. Also, at first it looks at distractors that are positioned on the same side of the room as the target but not being looked at by the caregiver, and at later stages corrects this [45]. The model also disregards eye direction cues in favor of head direction cues early on, but not so at later stages [46]. The model offers a possible explanation for the diminished gaze following in autistic individuals [47]: assigning a very small or even negative saliency to the caregiver considerably slows down or even abolishes the development of gaze following in the model. Thus, poor gaze following in autistic individuals may in part be caused by their aversion to social stimuli in general, and faces [48] and eye contact [49] in particular.

3.2 Integration of top-down attention

The model infant first learns to attend to salient objects (bottom-up attention). This is done by learning a one-to-one mapping between saliency detected at a particular location (which activates an element in $\mathbf{s}$) and the action of looking at that same location (the corresponding element in $\mathbf{m}$).

Gaze following (top-down attention) takes longer to learn because there is a one-to-many relationship between a caregiver looking direction (elements in $\mathbf{h}$ and $\mathbf{e}$) and the actions (elements in $\mathbf{m}$) corresponding to looking at locations along the caregiver's corresponding line of sight (see Fig. 3). Additionally, the model loses opportunities to learn to follow gaze in the times when the caregiver is not present.

To see how top-down visual search is gradually integrated in the model, an experimental setup was created, as depicted in Fig. 4: Trials start with the infant

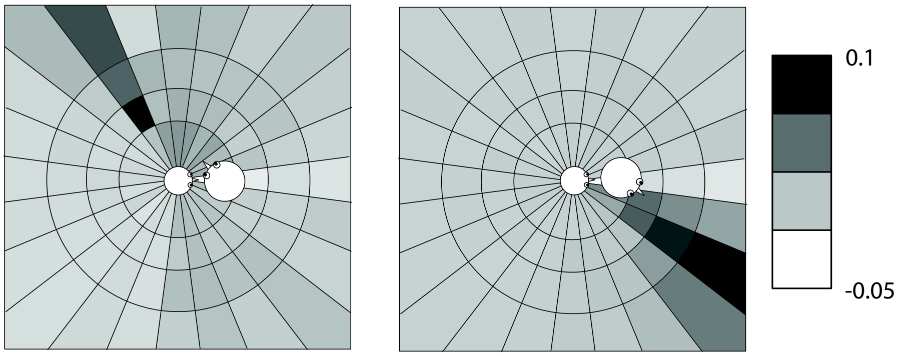

Fig. 3. Illustration of connection weights from inputs **h** and **e** to vector **m** after gaze following is learned. Shown are two different caregiver head/eye directions and the corresponding activations in **m** (since the caregiver's head and eyes are aligned in these examples, the values of **h** and **e** are the same; having them unaligned would result in a slightly more spreaded activation in **m**).

looking at the caregiver, and the caregiver looking to the left at 60° from her midline, towards an object (object A). Another object (object B) is positioned on the opposite of the room from object A. Object A's saliency is 80% of object B's. Trials last 6 seconds, after which it is noted what object the infant turns gaze to. If bottom-up influences are stronger than top-down influences, the infant will tend to look at object B, which is more salient but not being looked at. But as top-down influences are incorporated, the likelihood that the infant will disregard object B's saliency in favor of following the caregiver's gaze to object A will increase.

Trials were repeated 200 times, 100 for the setup shown in Fig. 4, and 100 for a "mirror setup", where objects A and B are swapped but with the caregiver still looking at object A. Fig. 5 shows the percentage trials in which the infant either looks at object A, object B, or at other (empty) locations. Before any learning could take place, at time step 0, the model's behavior corresponds to random action selection. Subsequently, as the value of bottom-up cues is learned, the infant preferably looks at object B, which is more salient. This reflects an increae in the connections of the bottom-up pathway from the saliency map to the pre-motor area. But as the infant later learns to follow gaze, it starts to look more at object A, which is less salient but being looked at by the caregiver. This shows a gradual integration of top-down attention into earlier bottom-up attention and corresponds to the development of appropriate connections between the representation of the caregiver's head and eye orientation and the pre-motor area for planning gaze shifts.

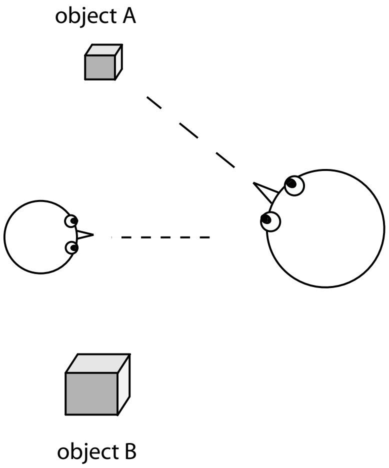

Fig. 4. Experimental setup for measuring bottom-up and top-down visual search. Object A is only 80% as salient as object B.

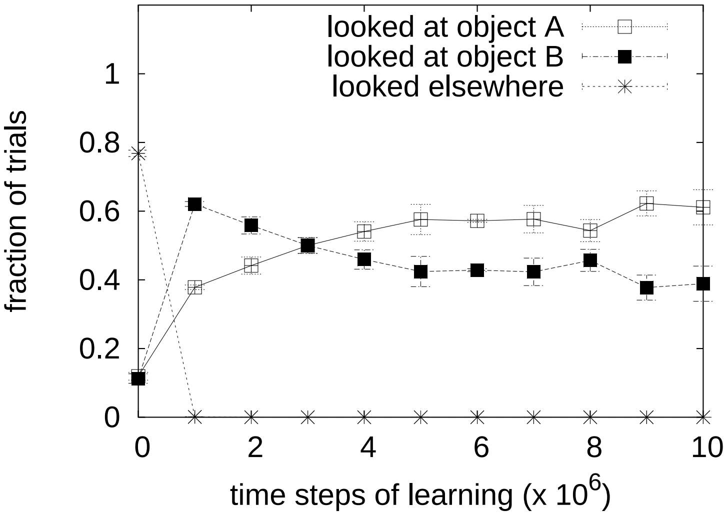

Fig. 5. Percentage trials where infant looks at object A, object B., or at other (empty) locations. From setup depicted in Fig. 4. Error bars indicate standard error after 5 repetitions.

4 Discussion

While much progress has been made in understanding the mechanisms and neural substrate of bottom-up attention [50, 51], our understanding of top-down attention is still in its infancy. Top-down control of visual attention occurs in a number of different ways including visual search, looking at locations in order to simplify motor control, prediction of when and where relevant information will be accessible, looking back to memorized locations, or attending to what other agents attend to. At present it is unclear, how these various mechanisms operate and how we, as infants, have acquired the ability to engage in these different forms of top-down attention control.

We have presented a model of the development of gaze following, that addresses how infants may learn to exploit the looking behavior of other agents to direct their attention to rewarding stimuli. Our model has a generic and biologically plausible reinforcement learning architecture. We have demonstrated how this model progressively incorporates top-down cues into its attention control system, learning to *optimally combine bottom-up and top-down processing pathways*, where optimality is defined in the sense of maximizing the obtained rewards. The model explains a large number of findings about the development of gaze following ability during infancy and makes a number of predictions. Most interestingly, maybe, it predicts the existence of a new class of mirror neurons specific for looking behaviors [39].

Technically speaking, there is not a big difference between the bottom-up and the top-down pathway of the model. In both cases, a number of “features” are mapped onto the same pre-motor representation via adjustable connection weights. For the bottom-up pathway, these features are perceived or remembered object saliencies in a body centered coordinate system. For the top-down pathway, these features represent the perceived or remembered orientation of another person’s head and eyes. The justification for refering to the latter pathway as “top-down” is that, according to our definition from above, it involves a more elaborate analysis of the visual scene beyond the mere calculation of a saliency map. The estimation of head pose and eye orientation from two-dimensional images is far from trivial and certainly requires a much more elaborate analysis than the computation of a saliency map. However, our model does not address how the infant may perform these computations, hence the distinction of the bottom-up and top-down pathways is admittedly only a formal one in our model.

It remains an open question, whether the development of other forms of top-down attention control can be understood in similar ways. We are optimistic that this may be the case and view reinforcement learning as a particularly useful perspective for understanding how other top-down attention mechanisms may be acquired, and how we learn to combine or integrate them with each other and with bottom-up mechanisms.

Acknowledgments

We are grateful to the members of the MESA (Modeling the Emergence of Shared Attention) project at UC San Diego. This work was supported by the National Science Foundation, a Marie Curie Excellence Center grant to J.T., and the Hertie Foundation.

References

1. Itti, L., & Koch, C. (2000). A saliency-based search mechanism for overt and covert shift of visual attention. *Vision Research*, *40*, 1489-1506.
2. Yarbus, A. L. (1967). Eye movements during perception of complex objects, in L. A. Riggs, ed., *Eye Movements and Vision*, Plenum Press, New York, chapter VII, pp. 171-196.
3. Land, M. F., Mennie, N. & Rusted, J. (1999). The roles of vision and eye movements in the control of activities of daily living. *Perception*, *28*, 1311-1328.
4. Land, M. F. & Hayhoe, M. M. (2001). In what ways do eye movements contribute to everyday activities? *Vision Research*, *41*, 3559-3565.
5. Hayhoe, M., Land, M., & Shrivastava, A. (1999). Coordination of eye and hand movements in a normal environment. *Invest. Ophthalmol & Vis. Sci.*, *40*, S380.
6. Wolfe, J. M. (1998). Visual search. In Pashler, H. (ed). *Attention*. London, UK: University College London Press.
7. Haith, M. M., Hazan, C., & Goodman, G. S. (1988). Expectation and anticipation of dynamic visual events by 3.5-month-old babies. *Child Development*, *59*, 467-479.
8. Dayan, P., & Abbott, L. F. (2001). *Theoretical Neuroscience: Computational and Mathematical Modeling of Neural Systems.* Cambridge, MA, USA: MIT Press.
9. Sutton R.S., & Barto, A.G. (1998). Reinforcement Learning. Cambridge, MA: MIT Press.
10. Scaife, M., & Bruner, J. (1975). The capacity for joint visual attention in the infant. *Nature*, *253*, 265-266.
11. Butterworth, G. E., & Cochran, E. (1980). Towards a mechanism of joint visual attention in human infancy. *International Journal of Behavioral Development*, *3*, 253272.
12. Butterworth, G. E., & Jarrett, N. (1991). What minds have in common is space: Spatial mechanisms serving joint visual attention in infancy. *British Journal of Developmental Psychology*, *9*, 5572.
13. Deák, G. , Flom, R. A., & Pick, A. D. (2000). Effects of Gesture and Target on 12- and 18-Month-Olds' Joint Visual Attention to Objects in Front of or Behind Them. *Developmental Psychology*, *36*, 511-523.
14. Brooks, R., & Meltzoff, A. N. (2002). The importance of eyes: How infants interpret adult looking behavior. *Developmental Psychology*, *38*, 958-966.
15. Brooks, R., & Meltzoff, A. N. (2005). The development of gaze following and its relation to language. *Developmental Science*, *8*, 535-543.
16. Meltzoff, A. N., & Brooks, R. (2007). Eyes wide shut: The importance of eyes in infant gaze following and understanding of other minds. In R. Flom, K. Lee & D. Muir (Eds.), *Gaze following: Its development and significance.* Mahwah, NJ: Erlbaum (pp.217-241).
17. Mundy, P., Sigman, M., & Kasari, C. (1990). A longitudinal study of joint attention and language development in autistic children. *Journal of Autism and Developmental Disorders*, *20*, 115-128.

18. Treisman, A., & Gelade, G. (1980). A feature integration theory of attention. *Cognitive Psychology*, *12*, 97-136.
19. Brooks, R., Breazeal, C., Marjanovic, M., Scassellati, B. and Williamson, M. The Cog Project: Building a Humanoid Robot. In C. Nehaniv, ed., Computation for Metaphors, Analogy and Agents, Vol. 1562 of Springer Lecture Notes in Artificial Intelligence, Springer-Verlag, 1998.
20. Kozima, H. and Yano, H. (2001). A robot that learns to communicate with human caregivers. In First International Workshop on Epigenetic Robotics (Lund, Sweden).
21. Kozima, H. (2002) Infanoid: A babybot that explores the social environment, K. Dautenhahn, A. H. Bond, L. Canamero, B. Edmonds (eds.), Socially Intelligent Agents: Creating Relationships with Computers and Robots, Amsterdam: Kluwer Academic Publishers, pp. 157-164.
22. Nagai, Y., Hosoda, K., Morita, A., & Asada, M. (2003). A constructive model for the development of joint attention. *Connection Science*, *15*, pp. 211-229.
23. Nagai, Y., Asada, M., & Hosoda, K. (2006) Learning for joint attention helped by functional development. Advanced Robotics, Vol. 20, No. 10, pp. 1165-1181, September 2006.
24. Hoffman, M. W., Grimes, D. B., Shon, A. P., Rao., R. P. N. (2006) A probabilistic model of gaze imitation and shared attention. *Neural Networks*, *19*, 299-310.
25. Matsuda G., & Omori T. (2001). Learning of Joint Visual Attention by Reinforcement Learning. In E. M. Altmann & A. Cleeremans (Eds.), *Proceedings of the fourth international conference on cognitive modeling* (pp. 157-162). Mahwah, NL: Lawrence Erlbaum Associates.
26. Triesch, J., Teuscher, C., Deák, G. O., & Carlson, E. (2006). Gaze following: why (not) learn it? *Developmental Science*, *9*, 125-157.
27. Lau, B. & Triesch, J. (2004). Learning gaze following in space: a computational model. *3rd International Conference for Development and Learning (ICDL'04)*, La Jolla, California, USA, October 20-22, 2004.
28. Daly, S., Matthews, K., & Ribas-Corbera, J. (1999) Visual eccentricity models in face-based video compression. *IS&SPIE Conference on Human Vision and Electronic Imaging IV*, San Jose, California, USA, January, 1999.
29. Stanley, J. C. (1976). Computer simulation of a model of habituation. *Nature*, *261*, 146-148.
30. Farroni, T., Csibra, G., Simion, F., Johnson, M. H. (2002) Eye contact detection in humans from birth. *Proc Natl Acad Sci U S A*, *99* pp. 96029605.
31. Li, Z. (1999). Contextual influences in V1 as a basis for pop out and asymmetry in visual search. In *Proceedings of the National Academy of Sciences*, *96*, 10530-10535.
32. Li, Z. (2002). A saliency map in primary visual cortex. *Trends in Cognitive Science*, *6*, 9-16.
33. von Hofsten, C., Dahlström, E., & Fredriksson, Y. (2005) 12-month-old infants' perception of attention direction in static video images. *Infancy*, *8*, 217-231.
34. Anstis, S. M., Mayhew, J. W., Morley, T. (1969) The perception of where a face or television 'portrait' is looking. *American Journal of Psychology*, *82*, 474-489.
35. Cline M. G. (1967) The perception of where a person is looking. *American Journal of Psychology*, *80*, 41-50.
36. Gibson, J. J., & Pick A. (1963) Perception of another person's looking behavior. *American Journal of Psychology*, *76*, 386-394.
37. Jenkins, R., Beaver, J. D., & Calder, A. J. (2006). I thought you were looking at me! Direction-specific aftereffects in gaze perception. *Psychological Science*, *17*, 506-514.

38. Schultz, W., Dayan, P., & Montague, P. R. (1997). A neural substrate of prediction and reward. *Nature*, *275*, 1593-1599.
39. Triesch, J., Jasso, H., & Deák, G. O. (2007). Emergency of Mirror Neurons in a Model of Gaze Following. *Adaptive Behavior, accepted.*
40. Fasel, I., Deák, G. O., Triesch, J., & Movellan, J. (2002). Combining Embodied Models and Empirical Research for Understanding the Development of Shared Attention. In *Proceedings of the International Conference on Development and Learning (ICDL02)*, Boston, MA, USA.
41. Johnson, M. H., Dziurawiec, S., Ellis, H. D. & Morton, J. (1991) Newborns' preferential tracking of face-like stimuli and its subsequent decline. *Cognition*, **40**, pp. 1-19.
42. Valenza, E., Simion, F., Cassia, V. M. & Umiltá, C. (1996) Face preference at birth. *J. Exp. Psychol. Hum. Percept. Perform*, **22**, pp. 892-903.
43. Murray, L., & Trevarthen, C. (1985). Emotion regulation of the interactions between two-month-olds and their mothers. In T. Field & N. Fox (Eds.), Social perception in infants (pp. 89-111). Norwood, NJ: Ablex.
44. Roberts, S., & Pashler, H. (2000) How persuasive is a good fit? A comment on theory testing. Psychological Review, vol. 107, pp. 358-367.
45. Jasso, H., Triesch, J., & Deák, G. O. (2006). A reinforcement learning model explains the development of gaze following. In *Proceedings of the 7th International Conference on Cognitive Modeling (ICCM 2006)*, Trieste, Italy.
46. Jasso, H. & Triesch, J. (2006). Using eye direction cues for gaze following - a developmental model. In *Proceedings of the 5th International Conference on Development and Learning (ICDL 2006)*, Bloomington, IN, USA.
47. Jasso, H. A reinforcement learning model of gaze following (2007). *Unpublished Ph.D. dissertation.* University of California, San Diego.
48. Adrien, J. L., Lenoir, P., Martineau, J., Perrot, A., Hameury, L., Larmande, C., & Sauvage, D. (1993). Blind ratings of early symptoms of autism based upon family home movies. *Journal of the American Academy of Child and Adolescent Psychiatry*, *32*, 617-626.
49. Hutt, C., & Ounsted, C. (1966). The biological signicance of gaze aversion with particular reference to the syndrome of infantile autism. *Behavioral Science*, 11 (5), 346-356.
50. Itti, L., & Koch, C. (2001). Computational modelling of visual attention. *Nature Reviews Neuroscience*, *2*, 194-203.
51. Zhaoping, L. (2002). A saliency map in primary visual cortex. *Trends in Cognitive Sciences 6*, 9-16