Learning gaze following in space: a computational model

Boris Lau[1,2] and Jochen Triesch[2]

[1]Department of Neuroinformatics
Ilmenau Technical University, Germany
boris.lau@stud.tu-ilmenau.de

[2]Department of Cognitive Science
UC San Diego, USA
triesch@cogsci.ucsd.edu

Abstract

Following another person's gaze in order to achieve joint attention is an important skill in human social interactions. This paper analyzes the gaze following problem and proposes a learning-based computational model for the emergence of gaze following skills in infants. The model acquires advanced gaze following skills by learning associations between caregiver head poses and positions in space, and utilizes depth perception to resolve spatial ambiguities.[*]

1 Introduction

1.1 Shared attention and gaze following

The capacity for shared attention or joint attention is a cornerstone of social intelligence. It refers to the matching of one's focus of attention with that of another person, which can be established for example by gaze following. The importance of attention sharing in infancy and early childhood is hard to overstate. It plays an important role in the communication between infant and caregiver [8]. It allows infants to learn what is important in their environment, based on the perceived "distribution of attention" of older, more expert individuals. In conjunction with a shared language, it makes children able to communicate about what they perceive and think about, and to construct mental representations of what others perceive and think about. Consequently, episodes of shared attention are crucial for language learning [13].

Some authors make a subtle distinction between joint and shared attention: Joint attention only requires that two individuals attend to the same object, whereas shared attention also implies that each have knowledge of the other individual's attention to this object. In this paper, we will only be concerned with joint visual attention, which has been defined as looking where somebody else is looking, and which we view as an important precursor to the emergence of true shared attention. While initially, joint visual attention is mostly initiated by the caregiver, young infants soon acquire gaze following skills and initiate joint attention themselves [2]. There has been a significant body of research studying how these skills develop since the pioneering work by Scaife and Bruner [10].

Two different kinds of theories of the emergence of gaze following have been proposed. The *modular or nativist theories* posit the existence of innate modules, which are typically thought to be the product of evolution rather than to emerge from learning (e.g. [1]). *Learning based accounts* explain the emergence of gaze following by postulating that infants learn that monitoring their caregiver's direction of gaze allows them to predict where interesting visual events occur. This idea goes back to Corkum & Moore [5]. At present, the experimental evidence for or against a learning account of the emergence of gaze following in infants is still inconclusive, but computational models have shown that it is possible to acquire gaze following skills through learning (see Sect. 2).

1.2 Developmental stages in gaze following

Different distinguishable stages and effects during the development of gaze following have been discovered in cross-sectional studies: Butterworth and Jarret tested gaze following abilities of 6-, 12- and 18-month-old infants in a controlled environment [3]. In their experiments the infants were seated facing their mothers at eye level in an undistracting laboratory. Two or four targets of identical shape and color were presented at the same time as pairs on opposite sides of the room, also at the infants' eye level. Mother and infant were facing each other in every trial, until the mother shifted her gaze to a designated target. The infants' reactions were monitored and analyzed. Figure 1 (left) shows a typical setup of the experiments. All tested infants could shift their gaze to the correct direction and were able to locate targets presented within their field of view. However, only the 18-month-old infants followed

[*]An earlier version of this paper has been presented at the workshop SOAVE2004 (**S**elf-**o**rganization of **a**dapti**v**e behavior), Ilmenau, Germany.

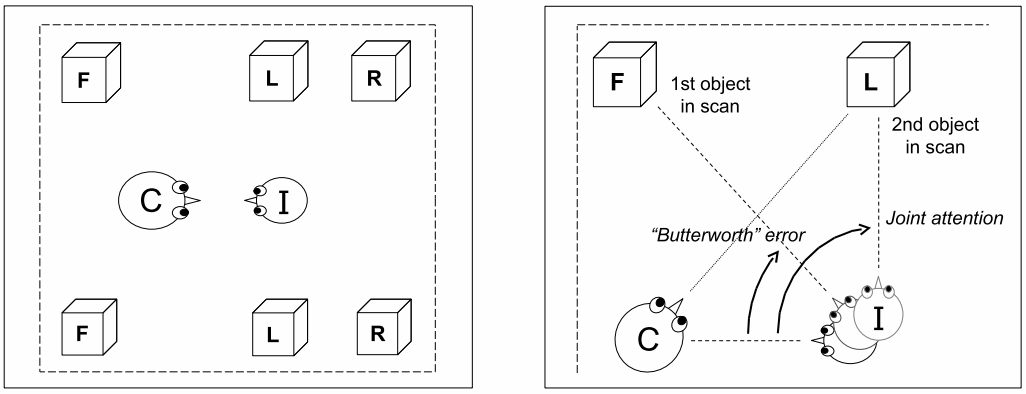

Figure 1. Left: Gaze following experiment with frontal (F), lateral (L) and rear (R) objects. Caregiver (C) and infant (I) are facing each other. Right: The caregiver looks at the lateral target. Six-month-old infants shift their gaze in the correct direction, but will most likely attend to the first object along their scan path (Butterworth error). 18-month-olds follow gaze to the correct lateral object, second in their scan path.

gaze to rear targets, while younger infants would not turn to search for targets behind them. When multiple target pairs were presented at the same time, for example the frontal and lateral targets in Fig. 1, 6-month-old infants were not able to tell which target their mother was looking at: when the mother turned to look at a lateral object, they shifted their gaze in the correct direction, but were likely to end the gaze shift at the first (frontal) object along their scan path, as shown in Fig. 1 (right). We call this effect the "Butterworth error". The infants in the 12 month group attended significantly more often to the correct object, but only the 18-month-old infants reliably followed their mother's gaze to the second (lateral) target.

Butterworth and Jarret associate a developmental stage with each of the age groups: Infants in the "ecological stage" around 6 months follow gaze in the right direction but locate only frontal targets correctly, and only if they are the first along the scan path. 12-month-old infants in the "geometric stage" are able locate the target objects more accurately and overcome the Butterworth error in some of the trials. Infants that have reached the "representational stage" around 18 months reliably overcome the Butterworth error and are also able to reliably locate targets behind them. The emergence of those stages is explained with three different mechanisms of gaze following that become effective in a sequential order and correspond to the observed stages [3].

1.3 Contribution of this paper

In order to explain the emergence of gaze following one has to explain the underlying dynamical processes of development, rather than just the snapshots provided by cross-sectional studies. In the remainder of this paper we will analyze the gaze following problem more carefully with an emphasis on its spatial properties, and isolate the different effects observed in the experimental studies. We propose a computational model, in which the infant acquires sophisticated gaze following skills and is able to overcome the Butterworth error by utilizing depth perception. It shows that the observed behaviors can emerge from the same learning mechanism and thus provides a more parsimonious account for the emergence of gaze following than the three different mechanisms proposed by Butterworth and Jarrett.

2 Gaze following and computational models

Following somebody's gaze in order to establish joint attention is a non-trivial task in cluttered environments. By observing someone's head pose, one can only infer the person's direction of gaze, rather than the distinct focus of the person's attention. Gaze following therefore requires scanning for an object along an estimate of a person's line of sight. For a precise estimate, infants have to evaluate the orientation of the caregiver's head and eye, as well as their own relative position to the caregiver. We will use the term 'head pose' in a general sense, referring to both head or eye orientations. The better the infants can discriminate different head poses, the better they can narrow down the region in space where they expect the caregiver's gaze target to be. Accurate depth perception can help to judge if objects are in the estimated line of gaze, and seems to be critical in situations where objects are in the projection of the caregiver's line of gaze but at different distances, as in Butterworth's

experiments. There is evidence that infants' perception of some depth cues continues to develop until at least 7 months [14]. This could have an impact on infants' ability to acquire advanced gaze following skills and may be part of an explanation of the staged development of gaze following.

We believe that infants typically learn the ambiguous mapping from caregiver head poses to locations in space without explicit supervision. Our goal is to plausibly explain this learning process by developing computer models that show how these skills can be acquired. In general, computational models have been developed that address different aspects of the gaze following problem. To our knowledge, two of them show how infants can learn gaze following without external task evaluation (no special reward for establishing joint attention) in a self-organizing manner. Both are discussed in the remainder of this section.

Carlson and Triesch recently proposed a computational model for the emergence of gaze following [4]. Their model infant predicts where salient objects are on the basis of the caregiver's head pose. They use a temporal difference (TD) learning approach [11] to show how an infant can develop these skills only driven by visual reward. The infant receives different rewards for looking at the caregiver and looking at salient objects. This reward structure can be adjusted to simulate certain symptoms of developmental disabilities like Autism or Williams Syndrome. Experiments with the model make predictions of the emergence of gaze following in children with those disabilities. Further experiments with this model were conducted by Teuscher and Triesch [12], focusing on the effect of different caregiver behaviors on infants' gaze following skills.

The model operates on a finite set of possible object locations without any spatial relationships. Each location has a one-to-one correspondence with a distinct caregiver head pose. One object is located at any time at any one of these positions. The caregiver agent has a certain probability of looking at that object. The model infant consists of two reinforcement learning agents: The 'when-agent' decides whether to continue fixating on the same location or to shift gaze, while the 'where-agent' determines the target of each gaze shift. Both agents try to maximize the long term reward obtained by the infant. The infant perceives the caregiver's head pose whenever it attends to the caregiver, and learns to exploit the correlation between the head pose and the location of salient objects. This model supports the theory of the acquisition of gaze following by learning. However, it is not adequate for simulating or explaining the Butterworth stages since it does not deal with geometrical relationships and spatial ambiguities.

Nagai et al. proposed a model for an infant agent that has been implemented on a robot platform [9]. The robot learns to follow the gaze of a human caregiver by offline training with recorded examples. Two separate modules, one for visual attention and one for learning and evaluation, output motor commands for turning the robot's camera head. A probabilistic gate module decides which of the two proposed motor commands gets executed. The probability for selecting the output of the learning module is changed from zero to one according to a predefined sigmoid function during the learning process. The visual attention module locates faces and salient objects by extracting color, edge, motion, and face features from the camera images. It uses a visual feedback controller to shift the robot's attention towards interesting objects. The learning module consists of a three-layered neural network that learns a mapping from gray-level face images to motor commands by backpropagation. The network is trained with the current motor position as teacher signal and the caregiver image as input, whenever a salient object is fixated.

The authors mention that every head pose only specifies a line of gaze rather than a distinct location in space. They deal with this ambiguity by moving the cameras incrementally towards the learned coordinates and stopping the movement at the first encountered object. Their model does not include depth perception and cannot resolve situations where distracting objects lie in the projection of the caregiver's line of gaze in the camera images, but at a different distance (compare Fig. 1, right). The model is not able to overcome the Butterworth error, which seems to be an essential characteristic of geometrical gaze following skills in infants.

3 A model of gaze following in space

Our new model specifically addresses the spatial ambiguities in the learning process of gaze following, and is able to faithfully reproduce infants' abilities to resolve them. It consists of a simulated environment and two different agents, an infant (Inf) and its caregiver (CG). The infant learns to follow the caregiver's gaze by establishing associations between the caregiver's head pose and positions in space where interesting objects or events are likely to be present. This online learning mechanism is driven by visual feedback, based on the infant's preference to look at the caregiver's face and salient objects in its environment. The infant exploits the correlation between the caregiver's line of gaze and the locations of salient objects to learn associations between those two. The perceptual preferences and the ability to shift gaze to interesting objects are important prerequisites for the learning process, which we assume to begin before infants show simple gaze following behaviour (i.e. before an age of six months).

The environment is similar to the setups in the experiments by Butterworth and Jarrett [3], with both agents' eyes and all objects being at the same height from the floor. The learning process is divided into trials. Objects are placed at

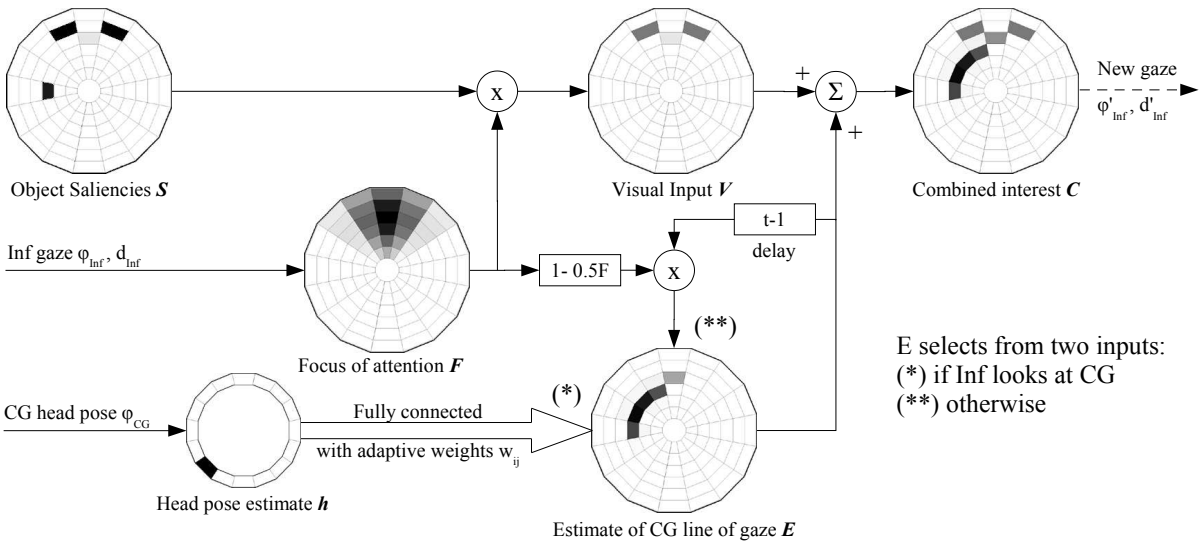

Figure 2. The infant agent with spatial representations in body-centered coordinate systems. Dark shading in the grid cells stands for high activation. The visual input V is the product of the object saliencies S and the focus of attention F. If the infant looks at the caregiver, the estimated caregiver head pose h is mapped to an estimate of the caregiver's line of gaze E. Otherwise the activation in E is held, inhibited by F to decrease the activation of locations along the line of gaze that the infant has already observed. E and V are summed up to the infant's combined interest in space C. The infant shifts its gaze to the area with the highest activation in C.

random positions in the environment in every trial. One of them is selected as the caregiver's focus of attention. The object locations and the caregiver direction of gaze do not change during a trial. The infant is looking at the caregiver at the beginning of every trial but can change its direction of gaze. The model operates on discrete time steps $t = 0, \ldots, T$. Each trial lasts for 10 time steps.

3.1 Environment, objects, and a caregiver

The environment is represented by a two-dimensional 7x9 grid with cartesian coordinates. Objects indexed with $i = 1, \ldots, N$ are introduced by specifying their grid coordinates (x_i, y_i) and a scalar saliency $s_i \in [0, 1]$. Both agents $a \in \{\text{Inf, CG}\}$ are defined by their positions in space (x_a, y_a), a base orientation φ_a^0 and the current direction of gaze $\varphi_a(t) \in [-180°, +180°]$, relative to φ_a^0. In addition to the current angle of gaze we introduce the function $d_a(t)$, which measures the distance from an agent to the point that the agent is currently looking at. The caregiver also has a saliency $s_{\text{CG}} = 0.1$. All angles and distances are discretized. We use 16 different values for angles (each corresponds to a range of $22.5°$), and 6 different values for distances (covering all possible distances in the 7x9 grid).

Since we focus on the spatial aspects of the learning problem and the infant's ability to learn gaze following without external task evaluation, we use a simple caregiver agent that does not react to the infant's actions. In every learning trial we let the caregiver look at the object i with the highest saliency s_i by setting its head/eye rotation $\varphi_{\text{CG}}(t)$ to the appropriate value.

3.2 The infant agent

The infant has to use its limited visual perception to gain information about the environment. The architecture of the infant agent is shown in Figure 2. It consists of different layers of neurons: the visual input V, the estimate of the CG line of gaze E, the combined interest C and the encoded caregiver head pose h. Their activations are represented with scalar values, assigned to the grid cells of a body-centered polar coordinate grid with discretized angle θ and radius r. The connections between those layers link only neurons encoding the same area in space. The object saliencies S and the encoded focus of attention F are also represented in body-centered coordinates. The infant's interest in the different locations in space is encoded by the combined interest layer $C(\theta, r, t)$. The activation of C is the sum of the visual input V and the estimate of the caregiver's line of gaze E:

$$C(\theta, r, t) := V(\theta, r, t) + E(\theta, r, t). \quad (1)$$

The infant shifts its gaze in every time step t to the area in space it is most interested in. This is done by setting its gaze orientation $\varphi_{\text{Inf}}(t)$ and looking distance $d(t)$ to the coordinates θ and r with the highest activation in $C(\theta, r, t)$.

Visual Perception is the infant's only source of information about its environment. It receives two different kinds of visual data: The caregiver head pose, encoded in the layer $h(\theta, t)$, and the actual visual input $V(\theta, r, t)$, which is the foveated transformation of the object's saliencies into the discretized polar coordinate system. V is used as a gate in the learning mechanism.

Generally we use discrete gaussian distributions $G_\sigma(x)$ as tuning curves for encoding input data for the infant agent. Extra normalization is necessary to ensure that the sum of the discrete distributions over all integers z is equal to one:

$$\tilde{G}_\sigma(x) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left(-\frac{x^2}{2\sigma^2}\right) \quad (2)$$

$$G_\sigma(x) = \tilde{G}_\sigma(x) / \sum_{z \in \mathcal{Z}} \tilde{G}_\sigma(z). \quad (3)$$

The caregiver's head pose $\tilde{\varphi}_{\text{CG}}$ is encoded with a population of neurons h with gaussian tuning curves. The variance ${\sigma_h}^2$ models the level of accuracy in head pose discrimination:

$$h(\theta, t) := G_{\sigma_h}\left(\varphi_{\text{CG}}(t) - \theta\right). \quad (4)$$

The locations (x, y) of the salient objects and caregiver are expressed in the infant's body-centered polar coordinates (θ', r'). The saliency value for each grid cell in S is the sum of all saliencies $s_k,\ k \in \{1, ..N, \text{CG}\}$ falling into the particular area of space. The infant's accuracy in depth perception is modeled with the variance ${\sigma_d}^2$ of the tuning curve encoding the distance of the objects:

$$S(\theta, r, t) := \sum_{k \mid \theta'_k = \theta} s_k(t) \cdot G_{\sigma_d}\left(r'_k(t) - r\right). \quad (5)$$

The infant's visual input V is the product of the object saliencies S and the focus of attention F, which is encoded in the same body centered coordinate system as the neural layers. It is a product of two gaussians (not normalized). It has its highest value at the focused point in the center of gaze $\theta = \varphi_{\text{Inf}}(t),\ r = d_{\text{Inf}}(t)$ and values close to zero for angles and distances further away from the infant's current focus of attention. This causes V to be a foveated view. The variances ${\sigma_\theta}^2$ and ${\sigma_r}^2$ influence the sharpness of the foveation:

$$V(\theta, r, t) := S(\theta, r, t) \cdot F(\theta, r, t) \quad (6)$$

$$F(\theta, r, t) := e^{-\frac{(\theta - \varphi_{\text{Inf}}(t))^2}{2{\sigma_r}^2}} \cdot e^{-\frac{(r - d_{\text{Inf}}(t))^2}{2{\sigma_\theta}^2}}. \quad (7)$$

Our model acquires gaze following skills by learning associations between the caregiver's head pose h and locations in space, forming the estimate of the caregiver's line of gaze $E(\theta, r, t)$. The associations are represented as connections with variable weights. We use a Hebbian-like learning rule that strengthens all connections from each active input neuron encoding a specific caregiver head pose to those locations where the infant saw a salient object shortly after observing the same head pose (activation in V). A small learning rate $\alpha_{\text{Hebb}} = 0.1$ combined with a slow decay of all synaptic weights, given by $\alpha_{\text{forget}} = 0.9999$, enables the network to 'forget' wrong associations that could be established when multiple objects are present during the training. The synaptic weight between a neuron j with activation $h(\omega, t)$ and a neuron i with activation $E(\theta, r)$ is given by $w_{ij}(t)$ and adapted with the following learning rule:

$$w_{ij}(t+1) := \alpha_{\text{forget}} \cdot w_{ij}(t) + \alpha_{\text{Hebb}} \cdot h(\omega, t) \cdot V(\theta, r, t). \quad (8)$$

The activation associated with the head pose encoded in h overwrites the activity in E whenever the infant is looking at the caregiver. When the infant has shifted its gaze away from the caregiver, E keeps its activation and is used as a short-term memory: the activation of the neurons encoding areas in space that the infant has already observed is suppressed by the activations of the neurons in F, encoding the focus of attention:

$$E(\theta, r, t) := \begin{cases} \sum_j \left\{w_{ij}(t) \cdot h(\omega, t)\right\}, \text{ if Inf looks at CG,} \\ E(\theta, r, t-1) \cdot \left(1 - \frac{1}{2} F(\theta, r, t)\right) \text{ otherwise.} \end{cases}$$

The selective inhibition of activity in E causes the infant to shift its gaze to unobserved locations, because it always attends to the area with the highest activation in C. This "scanning" continues as long as the activation along the line of gaze is higher than the activation due to the foveated visual input. It usually ends when the infant looks directly at an object.

4 Experiments

We present a number of experiments to show that our model infant is able to acquire gaze following skills and learns to overcome the Butterworth error. Each experiment is run 20 times under the same conditions for 1000 learning trials. The performance is measured in testing periods interposed every 25 trials during which no learning takes place. Every testing period consists of several trials with 10 time steps each, one trial for every tested object location. A trial is considered successful when the infant is looking where the caregiver is looking at the last time step of the trial. The performance of the model is measured with the

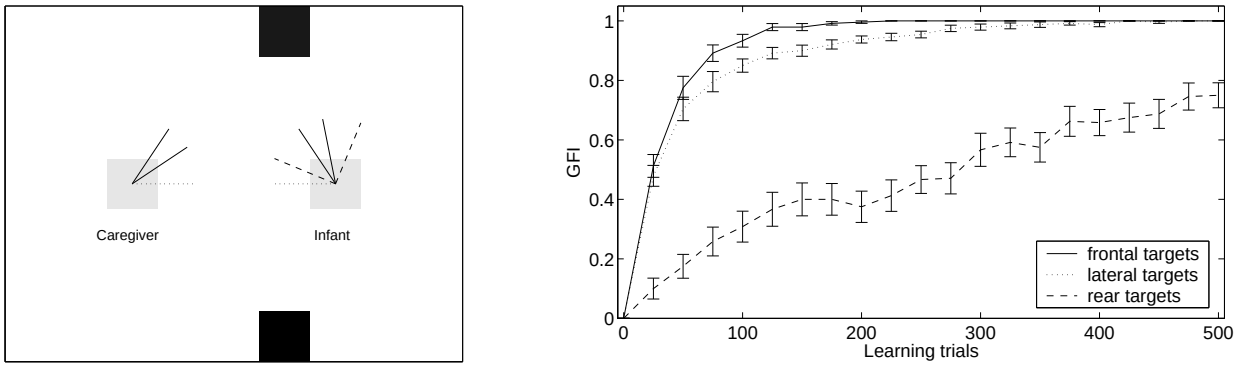

Figure 3. Gaze following performance for frontal, lateral and rear targets. *Left:* **Geometrical setup and situation at the end of a successful trial. The individual directions of the gaze of infant and caregiver are displayed with pairs of solid lines. The dotted lines indicate the agent's base orientation. The dashed lines display the borders of the infant's field of view.** *Right:* **Gaze Following Index for frontal, lateral and rear target pairs as functions of learning trials. The infant quickly learns to follow gaze to frontal and lateral targets. Gaze following to rear targets is acquired slowly. Data points are averaged from 20 runs, the error bars indicate the standard error.**

Gaze Following Index (GFI), which is defined as the number of successful trials divided by the total number of trials.

4.1 Gaze following performance

This experiment is designed to measure the model infant's gaze following performance separately for frontal, lateral and rear targets. We therefore split the testing trials in three groups, depending on the position of the caregiver's target object relative to the infant: a trial is considered a front target trial, when the caregiver's target is in the infants field of view while watching the caregiver. When the target object is initially out of view but not behind the infant, this is considered a lateral target trial. All other conditions are rear target trials.

Even the untrained model infant is able to locate frontal targets and to attend to them by simply using its peripheral vision. In order to eliminate this influence of simple preferential looking on the gaze following performance we present pairs of targets with a small difference in their saliency during the testing trials. Different from the learning trials we constrain the caregiver to look at the slightly less salient object in the testing trials, just by setting its head/eye rotation $\varphi_{\mathrm{CG}}(t)$ to the appropriate value. The infant will turn to the other, more salient object unless it follows the caregiver's gaze.

All individual target positions in space are tested, except the line connecting infant and caregiver. The setup is shown in Fig. 3 (left). We use tuning curves with small variances for encoding the caregiver head pose and the infant's perception of distances ($\sigma_h = \sigma_d = 0.1$) in order to test the gaze following performance independent from limitations in depth perception or face processing.

The result of this experiment is displayed in Fig. 3 (right). The infant learns to reliably follow the caregiver's gaze to frontal objects in about 100 learning trials, to lateral objects in about 200 learning trials, and to rear targets (with a little lower GFI) in about 500 trials. This corresponds to the results of the experiments by Butterworth and Jarret, where only the infants in the oldest age group shifted their gaze to rear targets.

The infant has not necessarily learned the complete set of associations for the frontal targets and every caregiver head pose until trial number 100. In fact, turning the head in the correct direction moves the target object closer to the infant's focus of attention and the other one further away. This can cause a higher activation in the foveated visual perception for the correct object than for the originally more salient distractor. In this case the infant will attend to the correct object. This corresponds to the ecological stage in the development in real infants.

A similar effect is exploited when the infant learns associations between a head pose and rear objects, outside the infant's field of view: Turning in the correct direction brings lateral targets into the infant's field of view and enables the infant to learn the corresponding associations. Learning to follow the caregiver's gaze to objects that are behind the infant requires a prior ability to follow gaze to lateral targets. This explains why it takes longer for the infant to achieve reliable gaze following skills for rear targets as seen in real infants.

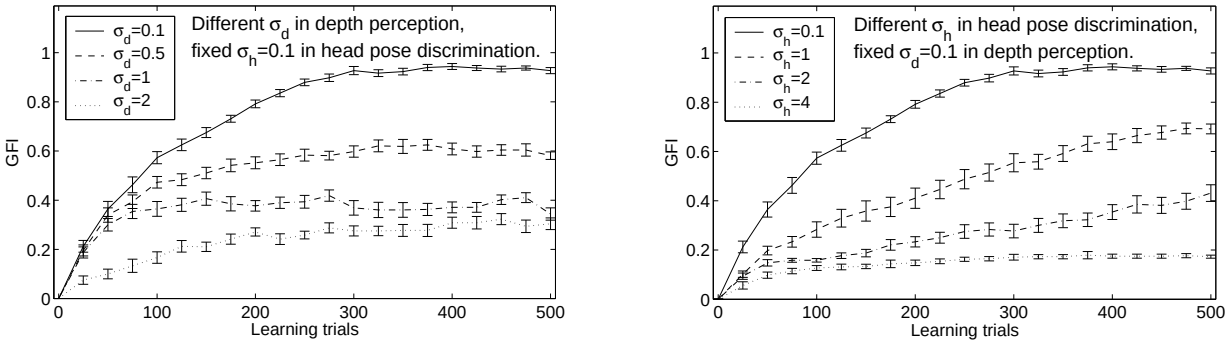

Figure 4. Overcoming the Butterworth error. Gaze Following Index for trials with lateral targets and frontal distractor objects, tested with different levels of accuracy in depth perception and head pose discrimination. High accuracy corresponds to using low variances for the tuning curves encoding object distances and caregiver head pose. Data points are averaged from 20 runs, the error bars indicate the standard error.

4.2 Overcoming the Butterworth error

In this experiment we test the infant's gaze following performance in the presence of distractor objects. Two salient distractors are placed as a pair of frontal targets behind the caregiver like shown in Fig. 1 (left). The slightly less salient target object, which the caregiver is attending to, is placed at different lateral and frontal locations, but not behind the infant or the caregiver. We test the gaze following performance with different settings for the infants ability to discriminate distances and head poses, by varying the variances ${\sigma_h}^2$ and ${\sigma_d}^2$ for the tuning curves encoding the head pose and the distances of the objects.

The results of this experiment are displayed in Fig. 4. The infant is able to overcome the Butterworth error and to ignore the distractor objects in the background for the majority of target positions, if depth perception and the discrimination of head poses are sufficiently accurate ($\sigma_h = \sigma_d = 0.1$). A higher variance (less accuracy) for depth perception or head pose discrimination leads to significantly worse gaze following performance. Unlike our model infant we assume real infants to improve their skills of depth perception and face processing over time. Our experimental results suggest that an infant cannot acquire geometrical gaze following skills before its depth perception and face processing skills are sufficiently developed. It is important to note that those skills seem to be critical not only for the actual gaze following, but for the acquisition as well.

Our model needs more than 200 learning trials to achieve reliable gaze following performance in the presence of distractors, compared to 100 trials in a simple setup with only one pair of objects. In both cases the model used high accuracy in depth perception and face processing from the first learning trial on. With only gradually developing depth perception skills the model would overcome the Butterworth error even later. These results correspond to the results of Butterworth where only older children are able to follow their caregiver's gaze correctly in ambiguous situations.

5 Discussion

We have analyzed the gaze following problem with an emphasis on its spatial characteristics, and presented a new model for the emergence of gaze following. The infant in our model learns to follow the caregiver's gaze by learning associations between observed head poses and positions in space. These associations form an ambiguous mapping from every head pose to several locations where salient objects are likely to be present. We demonstrated in experiments that our model is able to reach all stages of gaze following: first it is able to resolve spatial ambiguities when distractor objects are present in the background by using depth perception, and second it follows the caregiver's gaze to locations even behind its back. Furthermore, the temporal progression of the different stages is similar to the development observed in real infants: gaze following to frontal targets early in the development, overcoming the butterworth error and finding lateral targets later, and locating rear targets even later.

The model also makes predictions about the effect of limitations in depth perception and face processing on infants' ability to gain advanced gaze following skills: The better an infant can discriminate different head poses and object distances, the smaller is the region in space that will be associated with each head pose. If one of these two skills is not sufficiently developed, the model cannot overcome the Butterworth error. This suggests that children who are late to acquire accurate face processing and depth percep-

tion may develop geometric gaze following skills later than their peers.

Butterworth and Jarrett proposed that the development of a representation of space that contains infant, caregiver, and objects corresponds to the infants' ability to follow gaze to rear targets. The body-centered coordinate systems that we use in the infant agent provide such a spatial representation. The results of our first experiment show that gaze following to rear targets might occur later, even with such a representation of space already in place.

Our model, like most models, makes many abstractions and simplifications. While focusing on the spatial problems of gaze following we especially simplified the dynamic aspects in this problem by running the simulation in discrete trials. Different problems occur with a continuous time line in a dynamic environment: The longer the infant turns away from the caregiver, the more likely it is that the caregiver has already shifted its gaze again, causing a growing uncertainty in the infant's estimate of the caregiver head pose.

Popular approaches from the research areas of active vision and machine learning could be applied to the gaze following problem. One can understand the infant's search for salient targets as a state estimation process, based on limited observations of the real state, which is the actual distribution of salient objects in the room. Research on Partially Observable Markov Decision Processes (POMDPs) deals with the problem of decision making in environments with hidden states (e.g. [7]). Denzler and Brown developed an information theoretical approach to optimal sensor parameter selection in object recognition [6]. A similar approach could be used in the infant agent to learn how to efficiently integrate information from the available sources, namely accurate but visual perception with a limited field of view and ambiguous information from evaluating the caregiver's head pose.

Acknowledgments

The work described in this paper is part of the MESA project (**M**odeling the **E**mergence of **S**hared **A**ttention) at UC San Diego, a larger effort to understand the emergence of shared attention in normal and abnormal development supported by the National Alliance for Autism Research. We especially thank Christof Teuscher and Gedeon Deák for fruitful discussions, and Alan Robinson and Erik Murphy-Chutorian for comments on the draft. B. Lau is supported by the German National Merit Foundation.

References

[1] S. Baron-Cohen. *Mindblindness: an essay on autism and theory of mind*. A Bradford Book, The MIT Press, 1995.

[2] G. E. Butterworth. The ontogeny and phylogeny of joint visual attention. In A. Whiten, editor, *Natural theories of mind: Evolution, development, and simulation of everyday mindreading*, pages 223–232. Blackwell, 1991.

[3] G. E. Butterworth and N. Jarrett. What minds have in common in space: Spatial mechanisms serving joint visual attention in infancy. *British Journal of Developmental Psychology*, 9:55–72, 1991.

[4] E. Carlson and J. Triesch. A computational model of the emergence of gaze following. In H. Bowman and C. Labiouse, editors, *Connectionist Models of Cognition and Perception II*. World Scientific, 2003.

[5] V. Corkum and C. Moore. Development of joint visual attention in infants. In C. Moore and P. J. Dunham, editors, *Joint attention: Its origins and role in development*, pages 61–83. Erlbaum, 1995.

[6] J. Denzler and C. Brown. Information theoretic sensor data selection for active object recognition and state estimation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 24(2):145–157, 2002.

[7] L. P. Kaelbling, M. L. Littman, and A. R. Cassandra. Planning and acting in partially observable stochastic domains. *Artificial Intelligence*, 101:99–134, 1998.

[8] C. Moore and P. J. Dunham, editors. *Joint attention: Its origins and role in development*. Erlbaum, 1995.

[9] Y. Nagai, K. Hosoda, A. Morita, and M. Asada. A constructive model for the development of joint attention. *Connection Science*, 15(4):211–229, December 2003.

[10] M. Scaife and J. S. Bruner. The capacity for joint visual attention in the infant. *Nature*, 253:265ff, 1975.

[11] R. S. Sutton and A. G. Barto. *Reinforcement Learning: an introduction*. A Bradford Book, The MIT Press, 1998.

[12] C. Teuscher and J. Triesch. To care or not to care: Analyzing the caregiver in a computational gaze following framework. *3rd International Conference for Development and Learning, ICDL'04, La Jolla, California, USA*, 2004.

[13] M. Tomasello. *The cultural origins of human cognition*. Harvard Univ. Press, 1999.

[14] A. Yonas, C. A. Elieff, and M. E. Arterberry. Emergence of sensitivity to pictorial depth cues: Charting development in individual infants. *Infant Behaviour & Development*, 25:495–514, 2002.