Technische Universität Ilmenau
Fakultät für Informatik und Automatisierung
Fachgebiet Neuroinformatik und Kognitive Robotik

Learning gaze following in space: a computational model

Studienarbeit im Diplomstudiengang Informatik

Boris Lau

Entstanden während meines Aufenthalts im
Complex Systems & Cognition Laboratory
University of California, San Diego, USA

Betreuer:
Prof. Dr. Jochen Triesch
Department of Cognitive Science, UC San Diego, USA

Betreuender Hochschullehrer:
Prof. Dr. Horst-Michael Groß
Fachgebiet Neuroinformatik, TU Ilmenau, Deutschland

Eingereicht am 4. April 2006

Following another person's gaze in order to achieve joint attention is an important skill in human social interactions. This work analyzes geometric aspects of 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. It demonstrates that the succession of different "stages" of gaze following competence observed in infants can be explained with a single, generic learning mechanism.

Contents

0 Preface

The work for this paper has been done during my stay in the Complex Systems & Cognition Laboratory at the University of California, San Diego, with Jochen Triesch being my supervisor. It is part of the interdisciplinary 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. Parts of this paper have already been published and presented at the workshop "**S**elf**o**rganization of **a**dapti**ve** behaviour" in Ilmenau, Germany [1] and the "International Conference on Development and Learning" [2] in San Diego, USA. Compared to the published conference papers this seminar paper has a slightly longer introduction and an additional chapter about the new simulation platform, with implementation details of the model. Further work has been done on the model, to make it more robust with respect to parameter variations. It utilizes now a different way of representing memory, which I think is clearer and more straight forward. There also is a new section with a third experiment that tests how a cluttered training environment affects the performance of the model.

This paper could not exist without the help and support from several people: I want to thank Jochen Triesch, Christof Teuscher and Gedeon Deák for fruitful discussions, and Alan Robinson, Erik Murphy-Chutorian and Melanie Keller for comments on the draft. I also thank the German National Merit Foundation for their support, especially for the financial aid for my stay in San Diego.

1 Introduction

Why a computational model for the acquisition of gaze following in infants? Current accounts of cognitive development are largely descriptive. Little is known about how the many factors that change in the infant brain cause social specific skills to emerge. The developmental trajectory for a particular cognitive skill such as gaze following may depend in complex ways on the course of development of other skills such as face processing. Little is known about such dependencies and they are notoriously difficult to study experimentally. Computational models can help to theorize about developmental phenomena like the emergence of shared attention, and suggest explanations that in turn can guide further experimental work. The benefits of such an approach have been discussed in the literature (e.g. [3, 4]) and discovered by a growing community of scientists. Also, there are several international conferences that explicitly deal with cognitive modelling.

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. Attention sharing plays an important role in the communication between infant and caregiver [5]. 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 [6].

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

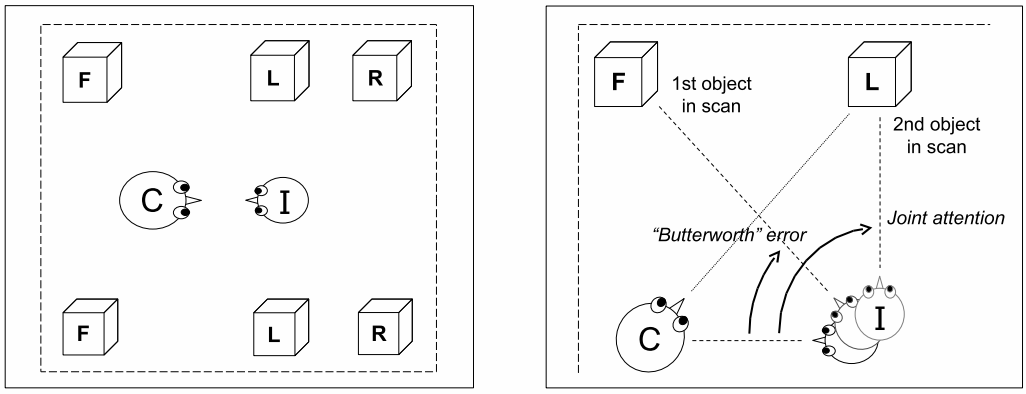

Figure 1.1: Left: Gaze following experiment with frontal (F), lateral (L) and rear (R) objects. Caregiver (C) and infant (I) are facing each other. Right: The caregiver looks at the lateral target. Six-month-old infants shift their gaze in the correct direction, but will most likely attend to the first object along their scan path (Butterworth error). 18-month-olds follow gaze to the correct lateral object, second in their scan path.

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 [7]. There has been a significant body of research studying how these skills develop since the pioneering work by Scaife and Bruner [8].

1.2 Developmental stages in gaze following

Infants do not follow another person's gaze from their birth on: three-month-old infants respond slowly or almost not at all to a parent's voice or gesture, but by 10 months, their responses are better organized, more controlled, and predictable. By their first birthday most infants can follow adults' gaze and pointing gestures [9, 10]. This schedule should be understood as a rough guideline. In fact, there has been a considerable debate about when gaze following emerges in human infants with estimates ranging from 6 to 18 months, and there is evidence for substantial individual differences [11].

Different distinguishable stages and effects during the development of gaze following

have been discovered in cross-sectional studies: Butterworth and Jarrett tested gaze following abilities of 6-, 12- and 18-month-old infants in a controlled environment [12]. 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 1.1 (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 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. 1.1, 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. 1.1 (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 Jarrett 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 first along the scan path. 12-month-old infants in the "geometric stage" are able to locate the target objects more accurately and overcome the Butterworth error in some of the trials. Infants who have reached the "representational stage" around 18 months overcome the Butterworth error even more reliable and are also able to 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 [12].

The postulation of three different mechanisms is not necessary to account for the observed patterns of behavior, however. Let us take a closer look at the geometry of gaze following. 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.

An important line of research is concerned with the specific features that infants use to estimate the adult's direction of gaze. There is evidence that younger infants rely exclusively on the caregiver's head pose [13], while between 12 and 14 months there is a significant increase in the reliance on the eye orientation [14]. By 18 months, gaze following is reliably produced on the basis of eye movements alone [12]. The models described in this paper do not explicitly differentiate between head and eye orientations. We will therefore use the term 'head pose' in a general meaning, 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 [15]. Thus, limitations in both head pose discrimination and depth perception 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.

1.3 Contribution of this work

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 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 demonstrates that the observed behaviors can emerge from a single 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 Previous computational models

Two different kinds of theories of the emergence of gaze following have been proposed. *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. [16]). *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 Moore & Corkum [17]. 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.

Several computational models have been developed that address different aspects of the gaze following problem. Two of them show how infants can learn gaze following without an external supervisor giving rewards for accomplished joint attention. Both are discussed in the remainder of this section.

Carlson and Triesch recently proposed a computational model for the emergence of gaze following [18]. 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 [19] 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 about the emergence of gaze following in children with those disabilities. Further experiments with this model were conducted by Teuscher and Triesch [20], focusing on the effect of different caregiver behaviors on the emergence of infants' gaze following skills.

The Carlson and Triesch 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 through temporal difference learning. 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 explaining the stages observed by Butterworth and Jarrett since it does not deal with geometric relationships and spatial ambiguities.

A model by Nagai et al. has been implemented on a robotic platform [21]. 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. 1.1, right). The model is not able to overcome the Butterworth error, which seems to be an essential characteristic of advanced gaze following skills in infants.

In conclusion, none of the models in the literature can correctly capture how infants eventually learn advanced gaze following skills as observed in 18-month-old infants.

3 A new gaze following model

Our new model specifically addresses the spatial ambiguities in the learning 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 for looking 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 the two. The perceptual preferences and the ability to shift gaze to interesting objects are important prerequisites for the learning process. We assume that both are operating before infants show simple gaze following behavior (i.e., before an age of six months).

The environment is similar to the setups in the experiments by Butterworth and Jarrett [12], 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 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 in discrete time steps $t = 0, \ldots, T$. Each trial lasts for $T = 10$ 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 $s_i \in [0.5, 1]$. 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 an associated saliency $s_{\mathrm{CG}} = 0.1$. 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 $\varphi_{\mathrm{CG}}(t)$ 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 3.1. It can be divided into three parts: a vision system for the perception of objects and their saliencies, a system for determining the caregiver's head pose and estimating its line of sight, and a memory and action selection system for shifting gaze to potential object locations. Across these three parts, there are different layers of neurons: the visual input V, the encoded caregiver head pose h, an interest layer I, two memory layers M and M_{gate}, and an action layer A. Their activations are represented with scalar values. All layers use a body-centered polar coordinate system with discretized angle θ and radius r. The only exception is the representation of caregiver's head pose h, which contains a representation of the caregiver's head orientation. Connections between the layers using body-centered representations link only neurons encoding the same area of space. Finally, there is a map of the saliencies of objects in the environment S and a focus of attention F, which is used to model foveated vision with a limited field of view in the infant. S and F should not be thought of as layers of neurons, but they also use the same body-centered representation. The activations shown in Fig. 3.1 correspond to the state of the model shown in Fig. 5.3.

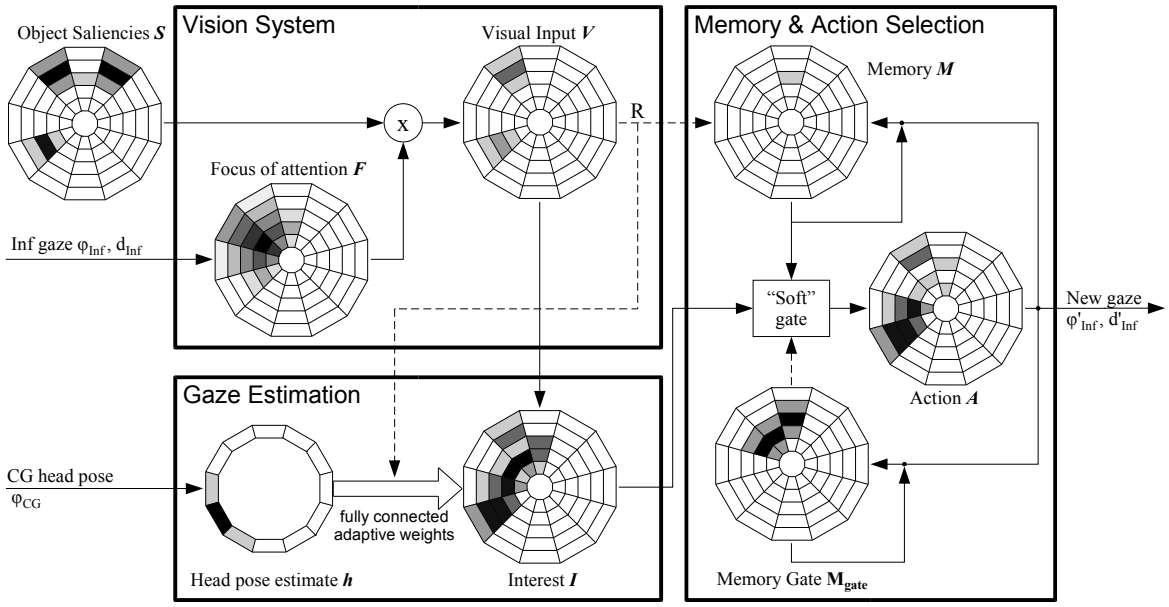

Figure 3.1: The infant agent with spatial representations in body-centered coordinate systems. Dark shading in the grid cells stands for high activation. The visual input V is the product of the object saliencies S and the focus of attention F. If the infant looks at the caregiver it estimates the caregiver's head pose h. This is mapped to an estimate of the caregiver's line of gaze, which is fused with the visual input into the interest map I. The infant shifts its gaze to the area with the highest activation in the action map A. M_{gate} is a memory representing which areas the infant has already inspected during the current trial, whereas M is a memory for the rewards received at these locations. Note that the discretization of space in the figure is coarser than the one actually used in the simulations.

3.2.1 Infant Vision System

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 $h(\theta, t)$ described below, and visually observed object saliencies, encoded in the visual input layer $V(\theta, r, t)$. In several places we use discretized probability density functions $G_\sigma(x)$ of the normal distribution as tuning curves for encoding input data for the infant agent. Extra normalization is necessary to ensure that the sum of the probabilities under the discrete gaussians over all integers x is equal to one. We define:

$$G_\sigma(x) = \frac{1}{Z} \exp\left(-\frac{x^2}{2\sigma^2}\right), \text{with } Z \text{ chosen such that } \sum_{x \in \mathcal{X}} G_\sigma(x) = 1\,. \tag{3.1}$$

The locations (x, y) of the salient objects and the caregiver are expressed as body-centered polar coordinates (θ', r'). The saliency value for each grid cell in S is the sum of all saliencies $s_k,\ k \in \{1, ..N, \mathrm{CG}\}$ falling into the particular area of space. The infant's limited accuracy in depth perception is modeled by a gaussian tuning curve that "blurs" the saliencies in S:

$$S(\theta, r, t) := \sum_{k \,|\, \theta'_k = \theta} s_k(t) \cdot G_{\sigma_d}\left(r'_k(t) - r\right). \tag{3.2}$$

The effect of the gaussian can be seen in the illustration of S in Fig. 3.1: There are three salient objects in the room (dark cells), but the infant is not absolutely sure about their distance. By changing the variance ${\sigma_d}^2$ of the gaussian we can control the infant's accuracy in depth perception.

We model the infant's vision as having a limited field of view and being foveated. Also, the infant's "depth of field" is limited, i.e. objects at distances other than the infant's current viewing distance are less salient than they would be at the right distance. To this end, we introduce the focus of attention F, which is encoded in the same body centered coordinate system as the neural layers. F is a product of two gaussians (not normalized):

$$F(\theta, r, t) := exp\left(-\frac{(\theta - \varphi_{\mathrm{Inf}}(t))^2}{2{\sigma_\theta}^2}\right) \cdot exp\left(-\frac{(r - d_{\mathrm{Inf}}(t))^2}{2{\sigma_r}^2}\right). \tag{3.3}$$

It has its highest value at the center of gaze $\theta = \varphi_{\mathrm{Inf}}(t),\ r = d_{\mathrm{Inf}}(t)$ and values close to zero for angles and distances further away from the infant's current center of gaze. The variances ${\sigma_\theta}^2$ and ${\sigma_r}^2$ influence the sharpness of the foveation and the size of the field of view. In this paper we use ${\sigma_\theta}^2 = 1$ and ${\sigma_r}^2 = 3.5$. The infant's visual input V is the product of the object saliencies S and the focus of attention F, which acts as a foveation function:

$$V(\theta, r, t) := S(\theta, r, t) \cdot F(\theta, r, t). \tag{3.4}$$

For locations far from the current center of gaze V is practically zero — these locations are outside of the infant's field of view.

In every time step the infant generates a scalar "reward" signal R from the visual input at the infant's center of gaze:

$$R(t) := V(\varphi_{\mathrm{Inf}}(t), d_{\mathrm{Inf}}(t))(t). \tag{3.5}$$

This reward is used as a gate in the learning mechanism for the connections between layers h and I described below.

3.2.2 Estimation of caregiver's line of gaze

The caregiver's head pose $\tilde{\varphi}_{\mathrm{CG}}$ 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. The activation in h is updated whenever the infant looks directly at the caregiver, using the following equation:

$$h(\omega, t) := G_{\sigma_h}\left(\varphi_{\mathrm{CG}}(t) - \omega\right). \tag{3.6}$$

The infant's interest in the different locations in space is encoded by the interest layer $I(\theta, r, t)$. The activation of I is the sum of the visual input V and the estimate of the caregiver's line of gaze, which is generated from the encoded caregiver head pose h. The neurons in h are fully connected to the neurons in I via adjustable weights. The activation in h is fed forward via these weights and added to the visual input:

$$I(\theta, r, t) := V(\theta, r, t) + \sum_{\omega} w_{\theta,r,\omega}(t) \cdot h(\omega, t)\ . \tag{3.7}$$

3.2.3 Memory and Action Selection

To prevent the infant from repeatedly checking the same location for interesting targets, we are introducing a simple memory mechanism, that allows the infant to keep track of which locations it has already looked at (layer M_{gate}) and what the observed saliencies at these locations were (layer M).

Without inaccuracy in depth perception, the perceived saliencies could simply be memorized by defining:

$$M(\varphi_{\mathrm{Inf}}, d_{\mathrm{Inf}})(t) := R(t)\ , \tag{3.8}$$

for every location inspected by the infant. Since the infant's depth perception can be inaccurate, a perceived saliency can be located at a distance not equal to $d_{\mathrm{Inf}}(t)$, and still cause a positive reward R for the distance $d_{\mathrm{Inf}}(t)$ because of the gaussian blurring in S. Thus we need to apply the same blurring for the memory of the perceived saliencies. This requires us to not only update the memory for the location inspected but also the memory for nearby locations — albeit to a lesser extent.

Concretely, we define for all r:

$$\begin{aligned} M(\varphi_{\text{Inf}}, r, t) := \quad & (1 - G_{\sigma_d}(d_{\text{Inf}}(t) - r)) \cdot M(\varphi_{\text{Inf}}, r, t-1) \\ & + G_{\sigma_d}(d_{\text{Inf}}(t) - r) \cdot R(t) \ . \end{aligned} \tag{3.9}$$

The parameter ${\sigma_d}^2$ modeling the spatial accuracy of the memory is identical to the one used in (3.2). The memory gate M_{gate} is used to keep track of the locations the infant has already looked at during this trial. Its activation is defined analogously to the one for M:

$$\begin{aligned} M_{\text{gate}}(\varphi_{\text{Inf}}, r, t) := \quad & (1 - G_{\sigma_d}(d_{\text{Inf}}(t) - r)) \cdot M_{\text{gate}}(\varphi_{\text{Inf}}, r, t-1) \\ & + G_{\sigma_d}(d_{\text{Inf}}(t) - r) \cdot 1 \ . \end{aligned} \tag{3.10}$$

The action map $A(\theta, r, t)$, finally, determines where the infant will look next. It is defined as:

$$A(\theta, r, t) := M_{\text{gate}}(\theta, r, t) \cdot M(\theta, r, t) + (1 - M_{\text{gate}}(\theta, r, t)) \cdot I(\theta, r, t) \ . \tag{3.11}$$

When a location has already been visited ($M_{\text{gate}}(\theta, r, t) > 0$), the memory of the previously observed reward stored in $M(\theta, r, t)$ will contribute to the activity in $A(\theta, r, t)$ at this location. If the location has not been visited before ($M_{\text{gate}}(\theta, r, t) = 0$), then the activity $A(\theta, r, t)$ is driven purely by the interest I for that location. Thus, M_{gate} functions as a "soft" gate mediating between the interest I and the memory of observed rewards M.

At every time step t the infant shifts its gaze to the area in space that corresponds to the highest activation in A. 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:

$$(\varphi_{Inf}(t+1), d_{Inf}(t+1)) := \arg\max_{(\theta, r)} A(\theta, r, t) \ . \tag{3.12}$$

To understand the interplay of the interest I on the one hand, and the memory and action selection on the other hand, it is best to look at an example. Consider the situation depicted in Figure 3.1. The infant has just looked at the caregiver and obtained the pose. Due to prior learning of the connections from the head pose to the caregiver's estimated line of gaze, there was a stronger activation in I where the

caregiver is potentially looking (the left side) compared to the right side. Consequently the infant subsequently made two small gaze shifts to the left. During these gaze shifts, however, no significant reward was observed because the distractor in the upper left is at the wrong depth. At this point, the memory gate indicates which positions have been looked at, the memory represents the observed saliencies at these locations. The memory of the poor reward at these locations prevents the infant from looking there again: the soft gating mechanism will discount the high interest I for these locations due to the low values in M. As a consequence, the infant will keep scanning in the same direction until it finds the target object in the lower left.

3.2.4 Learning

The 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 that is fed into $I(\theta, r, t)$. The associations are represented as connections with variable weights. We use a Hebbian 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 I). The reward $R(t)$ is used as a gate so that associations are only strengthened when the infant had really attended to a salient object. The synaptic weight between a neuron in the head pose representation with activation $h(\omega, t)$ and a neuron in the interest layer with activation $I(\theta, r)$ is given by $w_{\theta,r,\omega}(t)$ and adapted with the following learning rule:

$$w_{\theta,r,\omega}(t+1) := (1 - \alpha_{\text{forget}}) \cdot w_{\theta,r,\omega}(t) + \alpha_{\text{Hebb}} \cdot h(\omega, t) \cdot I(\theta, r, t) \cdot R(t) \; . \tag{3.13}$$

The gated Hebbian learning with learning rate $\alpha_{\text{Hebb}} = 0.015$ is combined with a slow decay of all synaptic weights, where we use $\alpha_{\text{forget}} = 5 \times 10^{-4}$. This enables the network to "forget" wrong associations that could be established when multiple objects are present during the training.

4 Our simulation environment

We have implemented the computational model presented in this paper completely in Matlab. It has been written from scratch, since the previous gaze following model did not include any spatial properties. The simulation environment consists of the model itself with an infant and a caregiver agent, scripts to generate the stimuli and to evaluate experiments. A set of visualization tools has been designed to complement the evaluation of numerical and statistical data. It consists of polar plots as used in Fig. 3.1, a three-dimensional display of the mapping matrix and a two-layer view that shows the interest I and the orientation of the agents in the room. The visualization was very important during the design process of the model to really see how the agents are reacting to the stimuli.

4.1 Implementation

During the development of this model we had to run it a lot of times, because we tested different model architectures with different parameter settings, each setting with 10 to 20 repetitions. The experiments also utilize parameter sweeps, with 20 runs per combination. Although not critical, a rather fast simulation of the model is therefore desirable. In the final version of the model the simulation takes roughly 20 seconds for one experimental run with 10^4 time steps, the first versions took more than 10 times as long. This improvement was achieved by optimizing the pure Matlab code, as we did not use any pre-compiled C-Functions. Matlab is comparatively fast when executing calculations on vectors or matrices, doing the same computation step by step in loops takes much longer. For functions that have to be called several times with different arguments we have used a little trick: All arguments are stored in a vector, which is then passed to the function in a single execution. This significantly shortened the computation of the gaussian functions, that we use all throughout our model to represent uncertainty. Examples for this can be found in the following code

Name	Value	Explanation
inf.visResolution(1)	16	#bins for discretizing angles
inf.visResolution(2)	6	#bins used for discretizing distances
inf.visRange(1)	$2pi$	angle that the infant can see with head rotation
inf.visRange(2)	6.7	euclidean distance from the infant to a far corner
inf.headPose(1)	φ_{Inf}	infant's current direction of gaze (bin number)
inf.headPose(2)	d_{Inf}	infant's current viewing distance (bin number)
fovHP, fovD	${\sigma_\theta}^2$, ${\sigma_r}^2$	Variances for sharpness of foveation
dNorm, hpNorm	Z	normalization factors for discrete gaussians
env		matrix representing the room with saliencies
maxObjects	N	number of random objects used during training

Table 4.1: Names of variables used in the code excerpts and their equivalent value or identifier in chapter 3, along with a short explanation.

excerpts. Table 4.1 lists a definition of the relevant variables that we use.

All discrete polar maps that we use to represent the neural layers in our model are stored in regular matrices, using the row index for the radius and the column index for the angle. For the matrices, gaussians are computed directly in the domain of row and column numbers. The mean of a gaussian thus is the indices of the matrix cell where it is centered on. In Matlab, indices for rows and columns always start with 1.

Coordinate transformations that are used regularly in the code of the model are implemented as seperate functions. Three different coordinate systems come to use here. First, cartesian coordinates (x, y) specify an absolute or relative position of an object or agent in the room. These coordinates directly index a cell in the matrix *env* that represents the room. Second, continuous polar coordinates (p, d) with a rotation angle p (pan)and a distance value d are used to specify the orientation and distance of an object or person relative to a person. The angle is measured in radiant and ranges from $-\pi$ to π, the distance is the euclidean distance taken from the cartesian space. Last, the discretized equivalents of the polar coordinates are used for the actual implementation of the polar maps. The discrete coordinates are natural numbers starting with 1 and index the discrete bins or cells in the matrices. For the discretization of the angles we want an orientation of 0° to fall into the "center" of a bin, rather then being on the border between two bins. In the conversion we therefore add 0.5 in the case there is an even total number of bins. A coordinate pair in this system is often called *hp* (for head pose) in the code excerpts.

The following two functions convert a relative room coordinate pair (x, y) to a (p, d)

pair, and a (p, d) pair to a head pose hp:

```
%%% convert relative room coordinates to polar coordinates
function x = xy2pd(pos)
  depth = sqrt(pos*pos'); % distance
  if depth==0 pan = 0;
  else
    if pos(2)>0 pan = -acos(pos(1)/depth);
    else pan = acos(pos(1)/depth);
    end
  end;
  x = [pan, depth];

%%% discretize pan and depth values (returning index numbers)
%   according to the persons visual resolution and range
function hp = pd2hp(pd, person)
  % make sure the pan angle is in -pi..pi
  if pd(1) > pi
    pd(1) = pd(1) - 2*pi;
  end
  if pd(1) < -pi
    pd(1) = pd(1) + 2*pi;
  end
  if (pd(2)>0) % if distance > 0
    % convert angle (in radiant) and distance (in gridcells) to bins
    hp1 = ((pd+[person.visRange(1)/2, 0])./(person.visRange));
    hp1 = hp .* (person.visResolution - [0 1]);
    % correction for an even total number of bins for pan
    hp1(1) = hp1(1) + 0.5*(mod(person.visResolution(1),2)==0);
    hp1(1) = mod(hp1(1), person.visResolution(1));
    % bring the values into 1..N range and discretize
    hp = floor(hp1+[1,1.499]);
  else
    hp = [0, 0];
  end
```

To obtain the focus of attention F according to (3.3) we initialize two arrays x_{hp} and x_d by assigning each cell its own index value. The size of these arrays are determined by the number of bins that are used for quantization. The two gaussians are generated for the index values, with the bin number for head pose and discretized viewing distance as mean. The function *gauss* is called with a mean, a variance and an argument vector as parameters, and returns the values of the gaussian for the given arguments. The product of the two gaussians yields F:

```
%%% compute the focus of attention F
% generate index arrays (x_hp(i)=i, x_d(i)=i)
x_hp = [1:inf.visResolution(1)]';
x_d  = [1:inf.visResolution(2)]';

% compute the gaussian functions
g_hp = gauss(inf.headPose(1), fovHP, x_hp);
g_d  = gauss(inf.headPose(2), fovD, x_d);
g_hp = g_hp;
g_d  = g_d;

F = g_hp*g_d';
```

To compute the polar map of object saliencies S perceived by the infant according to (3.2) we first obtain a list of all saliencies that are present in the room. For all these saliencies we transform the object's room coordinates X and Y to a continuous polar coordinate pair $polc$, specifying the direction $polc(1)$ and distance $polc(2)$ of the object relative to the infant. This pair is then discretized and transformed to a bin number pair hp, indexing the cell in the saliency matrix S that the saliency has to be assigned to. To account for uncertainty in the infant's depth perception we use a gaussian function that blurs the saliency in the matrix S. The object's distance $hp(2)$ is used as mean for the scaled gaussian that is added to the column of S that represent the direction of the object $hp(1)$ relative to the infant. The normalization factor $dNorm$ corresponds to a Z described in (3.1) to account for the discretization of the gaussian density function.

The infant's visual input V is the elementwise product of the saliencies and matrix representing the infant's focus of attention as in (3.4). The reward is computed exactly according to (3.5):

```
%%% get a map of the perceived saliencies
[X Y s] = find(env);
S = 0*S; % reset S
for i=1:size(X) % for all objects
  % coordinate transformations
  polc = xy2pd([X(i), Y(i)] - inf.pos) - inf.dir;
  hp = pd2hp(polc, inf);
  if hp>0
    % smudge for bad depth perception
    a = gauss(hp(2),dVariance,[1:person.visResolution(2)]') / dNorm;
    S(hp(1),:) = S(hp(1),:) + s(i)*a';
  end
end

V = F .* S;
R = V(inf.headPose(1), inf.headPose(2));
```

If the infant is looking at the caregiver, it renews its estimate of where the caregiver's line of gaze is. The estimated head pose is represented by an array with activation values for all discretized head poses. Again an index array is used to feed a gaussian functions with values. The bin number of the caregiver's head pose is used as mean, the variance defines the infant's accuracy in head pose discrimination. The result is an array with the blurred head pose estimate. Each activation value in the head pose estimate is now multiplied with the respective weight vector in the mapping matrix. The results are added up in the actHP matrix, which contains the estimate of the line of gaze. This matrix is added to the visual input to yield the infant's interest I:

```
%%% renew the gaze estimate
cgLoc = pd2hp(xy2pd(cg.pos-inf.pos)-inf.dir,inf);
if inf.headPose == cgLoc

  h = gauss(cg.headPose(1), hpVariance, [1:size(h,1)]');
  h = h / hpNorm;
  actHP = actHP*0; % reset the head pose estimate
  for i=1:size(h)
    actHP = actHP + w(:,:,i)*h(i);
```

```
  end
end

I = actHP + V;
```

The next code excerpt implements equations (3.9) to (3.12). The update of the polar memory and memory-gate maps according to (3.9) and (3.10) is based on the same gaussian g that models the distance component of the infant's focus of attention. For determining the infant's new head pose according to (3.12) we compute the maximum of the matrix A and search for cells in A with this value. Although the search could be stopped after one cell has matched the criterion, using the vectorized version of the find function is a lot faster. The indices of the first cell in the result yield the new head pose:

```
%%% memory and action selection
g = gauss(inf.headPose(2),dVariance,[1:inf.visResolution(2)]');
g = g / sum(g);
M(inf.headPose(1), :) = (1-g').*M(inf.headPose(1), :) + g'*R;
M_gate(inf.headPose(1), :) = (1-g').*M_gate(inf.headPose(1), :) + g';

A = M_gate.*M + (1-M_gate).*I;
% get the cell index with the highest activation in the action matrix
maxAct = max(max(max(A)));
[p,d] = find(A==maxAct);
% set the infant's head pose represented with discretized bin numbers
inf.headPose = [p(1),d(1)];
```

The learning of the mapping from head poses to regions in space happens through the adaptation of the weights w. The update is only done during the learn trials:

```
%%% adapt the weights
if learning
  for i=1:size(h)
    w(:,:,i) = (1-a_forget) * w(:,:,i) + h(i)*I*R*a_hebb;
  end
end
```

After each trial all activations in the infant are reset to zero, only the learned weights w stay the same. The infant's head pose is set such that the infant looks at the caregiver. One random object is generated and placed in the room in every learning trial. For test trials the objects are generated by the individual experiment scripts.

```
%%% initializes a new trial
actHP = actHP*0;
h = h*0;
I = I*0;
M_gate = M_gate*0;
M = M*0;
inf.headPose = cgLoc;

objects(:,3)=0;
env = env*0;
if learning
  for i=1:maxObjects
    % random x and y coordinates
    objects(i,1:2) = floor([rand, rand].*(roomsize))+1;
    % random saliency
    objects(i,3) = 0.5+0.5*rand;
    env(objects(i,1), objects(i,2)) = objects(i,3);
  end
end
```

After the initialization of a trial, the caregiver has to react and shift his gaze. For the regular learning trials the caregiver attends to the most salient object in the room. During testing trials we use a target with saliency 0.9 (see chapt. 5 for details):

```
%%% caregiver turns to the most salient object during learning,
% and to the target object with the saliency 0.9 during testing
if learning
  [x,y] = find(env==max(max(env)));
else
  [x,y] = find(env==0.9);
```

```
end
cg.target = [x(1),y(1)];
angle = xy2pd(cg.target - cg.pos) - cg.dir;
cg.headPose = pd2hp(angle, cg);
```

4.2 Modelling platforms for embodied models

The infant agent in our new model is in some ways an abstract form of an embodied model: it lives in an environment, senses and learns about its environment and acts according to its sensations and knowledge. On the other hand, there is no actual body of the model that has more characteristics than the position and orientation of the infant in the environment. Different platforms for the implementation of embodied models have been designed for research on developmental phenomena and, more specific, the emergence of shared attention.

Simulations in Virtual Reality (VR) allow to subtly make selective simplifications from real world scenarios, depending on the aspects of the model one wants to analyze. A good example for such a VR-platform has been developed in the scope of the MESA-project by our lab [22]. Robotic platforms have been designed to socially interact with humans. In the context of gaze following there have been attempts to let robots take the role of the infant or the role of the caregiver. Nagai et al. developed such a robotic child that learns to follow a human person's gaze (see chapt. 2). A research group at the UC San Diego constructed a robot that interacts with children and conducts for example gaze following tests [23]. A robotic head that can also be used as a platform for implementing developmental models has been developed in our lab [24].

5 Experiments

A number of experiments is presented to show that our model infant is able to acquire gaze following skills and learns to overcome the Butterworth error. We also demonstrate how a cluttered training environment affects the infant's gaze following performance. Each experiment is run 20 times under the same conditions for 1000 learning trials. The performance is measured in testing periods interposed every 50 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 Gaze Following Index (GFI), which is defined as the number of successful trials divided by the total number of trials.

5.1 Testing 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 frontal 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. In this experiment there is only one random object present during the training trials to provide the infant an "optimal" training environment.

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 two targets on opposite sides of the room with a small difference in their saliency during the testing

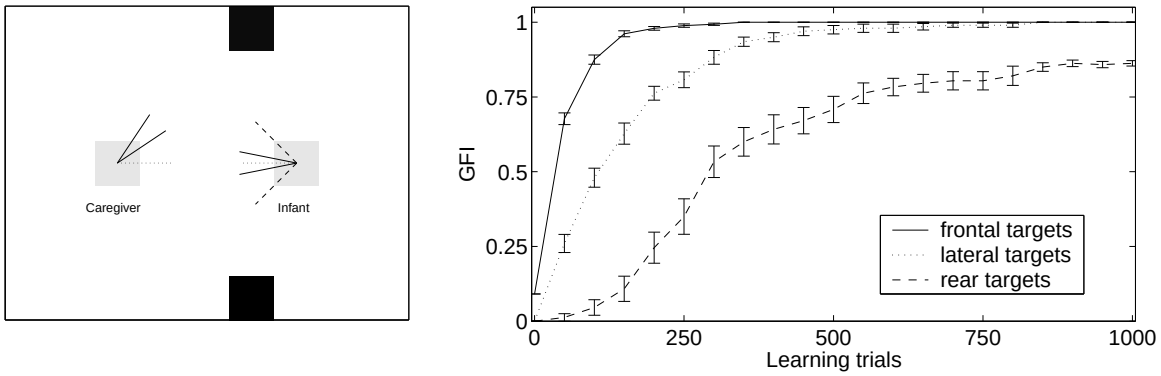

Figure 5.1: Gaze following performance for frontal, lateral and rear targets. *Left:* geometrical setup of a trial. The individual directions of the gaze of infant and caregiver are displayed with pairs of solid lines. The dotted lines indicate the agent's base orientation, the dashed lines display the borders of the infant's field of view. *Right:* Gaze Following Index for the different regions 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.

trials (we use 1.0 and 0.9 as saliencies). 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 reacts to the caregiver's direction of gaze.

All individual target positions in space are tested, except the line connecting infant and caregiver. The setup is shown in Fig. 5.1 (left). We use tuning curves with small variances for encoding the caregiver head pose and the infant's perception of distances (${\sigma_h}^2 = {\sigma_d}^2 = 0.1$) 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. 5.1 (right). Averaged for all target groups, our infant learns to reliably follow the caregiver's gaze (GFI > 0.75) in about 200 learning trials. Gaze following for frontal objects is learned in about 100 learning trials, to lateral objects in about 200 learning trials, and to rear targets in about 550 trials. This corresponds to the results of the experiments by Butterworth and Jarrett, where only the infants in the oldest age group shifted their gaze to rear targets.

Even though the infant attends to the correct target for nearly all frontal locations after 100 learning trials, it has not necessarily learned the complete set of associations for all positions at that time. A slight bias for shifting gaze to the correct side is

mostly sufficient to find the frontal targets: turning the head in the correct direction moves the infant's focus of attention (the maximum in F) closer to the target and further away from the other, originally more salient object, on the other side. This causes a higher activation in V and I for the cell representing the caregiver's target, and the infant will attend to this 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 than for frontal and lateral targets, as seen in real infants.

The results are robust with respect to scaling of the learning rates. If we increase or decrease both learning rates α_{Hebb} and α_{forget} by an order of magnitude while leaving their ratio constant, the results are qualitatively the same for learning to follow gaze to frontal and lateral targets, although the absolute number of trials that is necessary for the infant to achieve reliable gaze following will change. Gaze following to rear targets is more sensitive, since it requires more than just a small bias for turning into the correct direction to achieve gaze following.

5.2 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. 5.3 (left). The internal state of the infant agent in this situation can be seen in Fig. 3.1. The slightly less salient target object, which the caregiver is attending to, is placed at different lateral, frontal, and rear locations. 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. Since we want to isolate the effects that these two different limitations have on the infant's gaze following performance, we again provide an "optimal" training environment without clutter, and use only one random object during training.

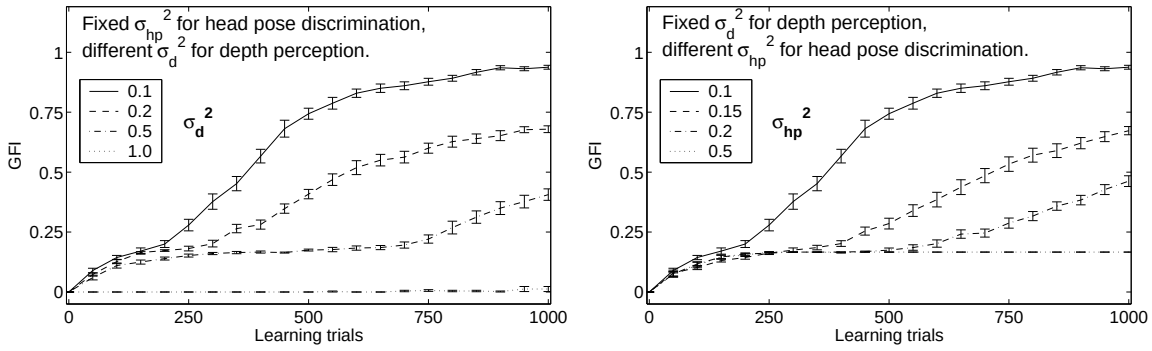

Figure 5.2: Overcoming the Butterworth error. Gaze Following Index for trials with two frontal distractor objects, tested with different levels of accuracy in depth perception and head pose discrimination. High accuracy corresponds to using low variances for the tuning curves encoding object distances and caregiver head pose. Data points are averaged from 20 runs, the error bars indicate the standard error.

The results of this experiment are displayed in Fig. 5.2. 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 (${\sigma_h}^2 = \sigma_d = 0.1^2$). 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 gradually improve their skills of depth perception and face processing over time. Unfortunately, very little is currently known about the exact time course of these processes, so we chose not to incorporate such a gradual improvement into our model. However, the present experimental results strongly suggest that an infant cannot acquire geometric 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, i.e., the infant can only *start* to learn advanced gaze following skills, when head pose discrimination and depth perception are sufficiently well developed.

Our model infant needs around 500 learning trials to achieve reliable gaze following performance in the presence of distractors, compared to 200 trials in the simple setup without distractors. 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

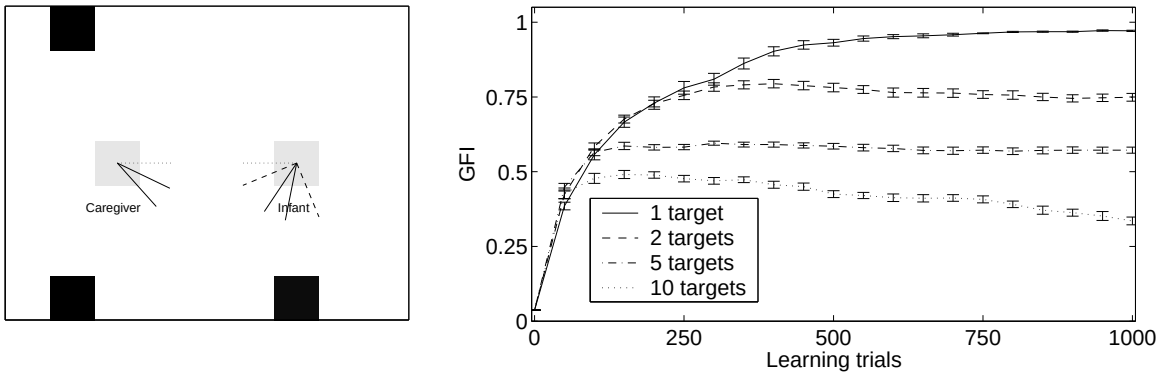

Figure 5.3: Left: Setup for experiment with distractor objects. The infant has already turned towards the target and has ignored the more salient distractors. Right: Gaze Following Index for testing trials with different numbers of random objects in the room during training. Data points are averaged from 20 runs, the error bars indicate the standard error.

able to follow their caregiver's gaze correctly in ambiguous situations. Note that in the experiments of Butterworth and Jarrett even 18-month-olds did not reliably follow gaze to rear targets in the presence of lateral distractors. The model eventually learns to do so, however. Thus, the model predicts that infants older than 18 month should eventually learn to follow gaze to rear targets in the presence of frontal distractors, too. To our knowledge, this experiment has not been attempted yet.

In this experiment the model is more sensitive to changes of the learning rates (only robust for scaling by up to a factor of 3, with slightly lower final performance): when the learning rate is too low, the distractor object causes a higher activity in I than the learned associations. On the other hand, a very high learning rate yields very high activations for the estimate of the caregiver's line of gaze in I, which will overpower the memorized rewards in the computation of A as in (3.11). Thus, the infant does not stop shifting gaze after finding the target object, but continues to scan for targets at all locations "slightly" associated with the caregiver's head pose. In this case, the infant might still return to the target after eliminating all other hypothesized target locations, but this would take significantly longer than the 10 time steps in the test trials, and is inconsistent with the behavior of real infants.

5.3 Learning in cluttered environments

In this experiment we test the robustness of the model with respect to the number of random objects present during the training trials. In the following we neglect the caregiver's own saliency, because it is significantly smaller than the saliency of the objects. If there is only one object present at a time, the caregiver always looks at this target. Since there are only "correct" stimuli, the infant associates the caregiver's head pose only with locations of objects that the caregiver has actually observed. With multiple targets being present during training the caregiver still reliably looks at the most salient object, but due to the foveated vision the infant will often shift its gaze to one of the other random objects, if the caregiver's more salient target is outside its field of view. This way, the infant also associates locations with the caregiver head pose that do not correspond to it. However, there is still a correlation between the caregiver's head pose and the location of salient objects, and we expect the infant to acquire gaze following skills at least to some degree.

Like in the first experiment, we test the gaze following performance with target pairs on all possible locations. Again, the caregiver turns to the slightly less salient object. This time there is no seperation into different target groups.

The result of this experiment is displayed in Fig. 5.3 (right). As expected, the more targets are present during training, the lower is the gaze following index that the infant achieves after 1000 learning trials. But even with a number of 10 objects, which can cover about 15% of the possible 9x7 locations in the room, the infant still shows gaze following behavior. This demonstrates that it is possible for the infant to acquire gaze following skills based on the correlation between head poses and locations of objects, even if not all training stimuli are "correct". On the other hand, in cluttered training environments the gaze following performance is significantly worse than it could be. Real infants learn gaze following in their everyday life, and their environment contains a lot more interesting things to see than just a few designated targets. In many cases, a child's home is cluttered with salient objects in many different colors. However, motion can let certain things like people, cars or moving toys stick out from the rest. Also, in social interactions, e.g. when playing with their children, caregivers tend to draw their infants' attention to the objects they currently look at, for example by slightly shaking a toy or moving their hands. It seems likely that this makes these targets temporarily more salient than other objects in the environment, thus relatively thinning out the clutter and enhancing the training conditions for the infant.

6 Conclusion

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, even in cluttered environments. These associations form an ambiguous mapping from every head pose to several locations where salient objects are likely to be present. Compared to a previous version of the model [2], we have only used strictly local learning rules in the current model. 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/or depth perception may develop geometric gaze following skills later than their peers. At present, only little is known about the head pose discrimination accuracy of infants in this age range and how it develops over time. Regarding depth perception, available evidence makes it seem likely that it was not the limiting factor in the experiments of Butterworth and Jarrett (e.g. [25], chap. 3).

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. Hence, it is premature to conclude that a lack of such a representation is responsible for failures to follow gaze to rear targets. While this is certainly a possibility, it is also conceivable that such a representation is already in place, but that it simply has not been properly *connected* to a representation of the caregiver's head pose.

Our model, like all models, makes many abstractions and simplifications. While focusing on the spatial problems of gaze following we especially simplified the dynamic aspects of the 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. The memory model for remembering the saliency of previously fixated locations suffers from a similar problem.

Our present model uses Hebbian learning. A re-formulation in the language of more modern learning approaches from the research areas of reinforcement learning, active vision or machine learning could be desirable. 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. [26]). Denzler and Brown developed an information theoretic approach to optimal sensor parameter selection in object recognition [27]. A similar approach could be used in the infant agent to learn how to efficiently integrate information from the available sources, namely accurate visual perception with a limited field of view and ambiguous information from evaluating the caregiver's head pose.

Bibliography

[1] B. Lau, J. Triesch, Learning gaze following in space: a computational model, in: Proc. 3rd Workshop on Self-Organization of AdaptiVE Behavior (SOAVE), Ilmenau, Germany, Fortschritt-Berichte VDI, Reihe 10, Nr. 743, VDI-Verlag, 2004.

[2] B. Lau, J. Triesch, Learning gaze following in space: a computational model, in: 3rd International Conference for Development and Learning, ICDL'04, La Jolla, California, USA, 2004.

[3] J. L. Elman, E. A. Bates, M. H. Johnson, D. Karmiloff-Smith, A.and Parisi, K. Plunkett, Rethinking Innateness: A connectionist perspective on development, A Bradford Book, The MIT Press, 1996.

[4] R. C. O'Reilly, Y. Munakata, Computational Explorations in Cognitive Neuroscience, A Bradford Book, The MIT Press, 2000.

[5] C. Moore, P. J. Dunham (Eds.), Joint attention: Its origins and role in development, Erlbaum, 1995.

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

[7] G. E. Butterworth, The ontogeny and phylogeny of joint visual attention, in: A. Whiten (Ed.), Natural theories of mind: Evolution, development, and simulation of everydaymindreading, Blackwell, 1991, pp. 223–232.

[8] M. Scaife, J. S. Bruner, The capacity for joint visual attention in the infant, Nature 253 (1975) 265ff.

[9] G. E. Butterworth, S. Itakura, How the eyes, head and hand serve definite reference, British Journal of Developmental Psychology 18 (2000) 25–50.

[10] G. O. Deák, R. Flom, A. D. Pick, Perceptual and motivational factors affecting joint visual attention in 12- and 18-month-olds, Developmental Psychology 36 (2000) 511–523.

[11] P. Mundy, A. Gomes, Individual differences in joint attention skill development in the secondyear, Infant Behaviour & Development 21 (2) (1998) 373–377.

[12] G. E. Butterworth, N. Jarrett, What minds have in common is space: Spatial mechanisms serving joint visualattention in infancy, British Journal of Developmental Psychology 9 (1991) 55–72.

[13] J. D. Lempers, Young children's production and comprehension of nonverbal deictic behaviors, Journal of Genetic Psychology 135 (1979) 93–102.

[14] A. J. Caron, S. C. Butler, R. Brooks, Gaze following at 12 and 14 months: Do the eyes matter?, British Journal of Developmental Psychology 20 (2002) 225–239.

[15] A. Yonas, C. A. Elieff, M. E. Arterberry, Emergence of sensitivity to pictorial depth cues: Charting development inindividual infants, Infant Behaviour & Development 25 (2002) 495–514.

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

[17] C. Moore, V. Corkum, Social understanding at the end of the first year of life, Developmental Review 14 (1994) 349–372.

[18] E. Carlson, J. Triesch, A computational model of the emergence of gaze following, in: H. Bowman, C. Labiouse (Eds.), Connectionist Models of Cognition and Perception II, World Scientific, 2003.

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

[20] C. Teuscher, 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.

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

[22] H. Jasso, J. Triesch, A virtual reality platform for studying cognitive development, in: 3rd International Conference for Development and Learning, ICDL'04, La Jolla, California, USA, 2004.

[23] J. R. Movellan, J. S. Watson, The development of gaze following as a bayesian systems identification problem, Tech. Rep. 2002.01, MPLab, UC San Diego, California, USA (2002).

[24] H. Kim, G. York, G. Burton, E. Murphy-Chutorian, J. Triesch, Design of an anthropomorphic robot head for studying autonomous development and learning, Proc. of IEEE 2004 International Conference on Robotics and Automation (ICRA 2004), Los Angeles, CA, USA.

[25] P. J. Kellman, M. E. Arterberry, The cradle of knowledge: development of perception in infancy, A Bradford Book, The MIT Press, 1998.

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

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