Sensory-motor Coordination in Gaze Control

G. de Croon, E.O. Postma, and H.J. van den Herik

IKAT, Universiteit Maastricht
P.O. Box 616, 6200 MD, Maastricht, The Netherlands
Voice: 0031-433883477 Fax: 0031-433884897
e-mail: g.decroon@cs.unimaas.nl, postma@cs.unimaas.nl, herik@cs.unimaas.nl
http://www.cs.unimaas.nl

Abstract. In the field of artificial intelligence, there is a considerable interest in the notion of sensory-motor coordination as an explanation for intelligent behaviour. However, there has been little research on sensory-motor coordination in tasks that go beyond low-level behavioural tasks. In this paper we show that sensory-motor coordination can also enhance performance on a high-level task: artificial gaze control for gender recognition in natural images. To investigate the advantage of sensory-motor coordination, we compare a non-situated model of gaze control (incapable of sensory-motor coordination) with a situated model of gaze control (capable of sensory-motor coordination). The non-situated model of gaze control shifts the gaze according to a fixed set of locations, optimised by an evolutionary algorithm. The situated model of gaze control determines gaze shifts on the basis of local inputs in a visual scene. An evolutionary algorithm optimises the model's gaze control policy. From the experiments performed, we may conclude that sensory-motor coordination contributes to artificial gaze control for the high-level task of gender recognition in natural images: the situated model outperforms the non-situated model. The mechanism of sensory-motor coordination establishes dependencies between multiple actions and observations that are exploited to optimise categorisation performance.

1 Introduction

In the field of artificial intelligence there is a considerable interest in situated models of intelligence that employ sensory-motor coordination to solve specific tasks [1, 2]. A situated model of intelligence is a model in which motor actions co-determine future sensory inputs. Together, the sensory inputs and the motor actions form a closed loop. Sensory-motor coordination exploits this closed loop in such a way that the performance on a particular task is optimised.

Several studies have investigated the mechanism of sensory-motor coordination [3–6]. For instance, they show that sensory-motor coordination can simplify the execution of tasks, so that the performance is enhanced. However, until now, research on sensory-motor coordination has only examined low-level tasks, e.g., categorising geometrical forms [3–7]. It is unknown to what extent sensory-motor coordination can contribute to high-level tasks.

So, the research question in this subdomain of AI research reads: *Can sensory-motor coordination contribute to performance of situated models on high-level tasks?* In this paper we restrict ourselves to the analysis of two models both performing the same task, viz. gaze control for gender recognition in natural images. The motivation for the choice of this task is two-fold: (1) it is a challenging task, to which no situated gaze control models have been applied before; (2) it enables the comparison of two models that are identical, except for their capability to coordinate sensory inputs and motor actions. Thus, we will compare a situated with a non-situated model of gaze control. If the situated model's performance is better, we focus on a second

research question: *How does the mechanism of sensory-motor coordination enhance the performance of the situated model on the task?* We explicitly state that we are interested in the relative performance of the models and the cause of an eventual difference in performance. It is not our intention to build the gender-recognition system with the best categorisation performance. Our only requirement is that the models perform above chance level (say 60% to 80%), so that a comparison is possible.

The rest of the paper is organised as follows. In Sect. 2 we describe the non-situated and the situated model of gaze control. In Sect. 3 we outline the experiment used to compare the two models of gaze control. In Sect. 4 we show the experimental results and analyse the gaze control policies involved. In Sect. 5 we discuss the relevance of the results. Finally, we draw our conclusions in Sect. 6.

2 Two Models of Gaze Control

Below, we describe the non-situated model of gaze control (Sect. 2.1) and the situated model of gaze control (Sect. 2.2). Then we discuss the adaptable parameters of both models (Sect. 2.3).

2.1 Non-situated Model of Gaze Control

The non-situated model consists of three modules. The first module receives the sensory input and extracts input features, given the current fixation location. The second module consists of a neural network that determines a categorisation based on the extracted input features. The third module controls the gaze shifts. Figure 1 shows an overview of the model. The three modules are illustrated by the dashed boxes, labelled 'I', 'II', and 'III'. The current fixation location is indicated by an 'x' in the face.

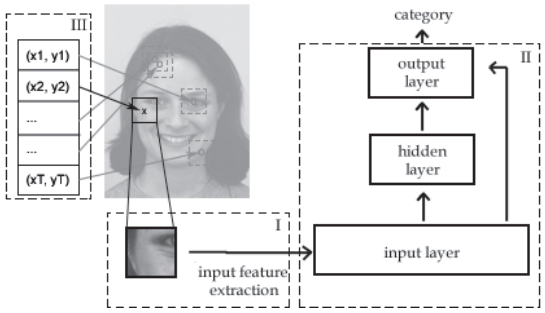

Fig. 1. Overview of the non-situated model of gaze control.

Module I receives the raw input from the window with centre x as sensory input. In Fig. 1 the raw input is shown on the left in box I; it contains a part of the face. From that window, input features are extracted (described later). These input features serve as input to module II, a neural network. The input layer of the neural network is illustrated by the box 'input layer'. Subsequently, the neural network calculates the activations of the hidden neurons in the 'hidden layer' and of the output neuron in the 'output layer'. There is one output neuron that indicates the category of the image. The third module determines the next fixation location, where the process is repeated.

Below we describe the three modules of the non-situated model of gaze control in more detail.

Module I: Sensory Input. Here, we focus on the extraction procedure of the input features. For our research, we adopt the set of input features as introduced in [8], but we apply them differently.

An input feature represents the difference in mean light intensity between two areas in the raw input window. These areas are determined by the feature's type and location. Figure 2 shows eight different types of input features (top row) and nine differently sized locations in the raw input window from which the input features can be extracted (middle row, left). The sizes vary from the whole raw input window to a quarter of the raw input window. In total, there are $8 \times 9 = 72$ different input features. In the figure, two example input features are given (middle row, right). Example feature 'L' is a combination of the first type and the second location, example feature 'R' of the third type and the sixth location. The bottom row of the figure illustrates how an input feature is calculated. We calculate an input feature by subtracting the mean light intensity in the image covered by the grey surface from the mean light intensity in the image covered by the white surface. The result is a real number in the interval $[-1, 1]$. In the case of example feature L, only the left half of the raw input window is involved in the calculation. The mean light intensity in the raw input window of area 'A' is subtracted from the mean light intensity of area 'B'.

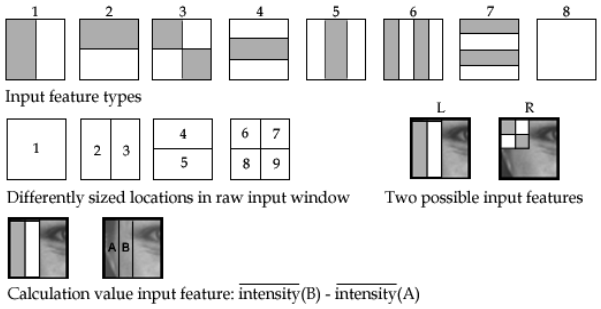

Fig. 2. An input feature consists of a type and a location.

Module II: Neural Network. The second module is a neural network that takes the extracted input features as inputs. It is a fully-connected feedforward neural network with h hidden neurons and one output neuron. The hidden and output neurons all have sigmoid activation functions: $a(x) = \tanh(x)$, $a(x) \in \langle -1, 1 \rangle$. The activation of the output neuron, o_1, determines the categorisation (c) as follows.

$$c = \begin{cases} \text{Male} & \text{, if } o_1 > 0 \\ \text{Female} & \text{, if } o_1 \leq 0 \end{cases} \tag{1}$$

Module III: Fixation locations. The third module controls the gaze in such a way, that for every image the same locations in the image are fixated. It contains coordinates that represent *all* locations that the non-situated model fixates. The model first shifts its gaze to location (x_1, y_1) and categorises the image. Then, it fixates the next location, (x_2, y_2), and again categorises the image. This process continues, so that the model fixates all locations from (x_1, y_1) to (x_T, y_T) in sequence, assigning a category to the image at every fixation. The performance is based on these categorisations (see Sect. 3.2). Out of all locations in an image, an evolutionary algorithm selects the T fixation locations. Selecting the fixation locations also implies selecting the order in which they are fixated.

2.2 Situated Model of Gaze Control

The situated model of gaze control (inspired by the model in [7]) is almost identical to the non-situated model of gaze control. The only difference is that the gaze shifts of the situated model are not determined by a third module, but by the neural network (Fig. 3). Therefore, the situated model has only two modules. Consequently, the current neural network has three output neurons. The first output neuron indicates the categorisation as in (1). The second and the third output neurons determine a gaze shift (Δx, Δy) as follows.

$$\Delta x = \lfloor m o_2 \rfloor \tag{2}$$

$$\Delta y = \lfloor m o_3 \rfloor, \tag{3}$$

where o_i, $i \in \{2, 3\}$, are the activations of the second and third output neurons. Moreover, m is the maximum number of pixels that the gaze can shift in the x- or y-direction. As a result, Δx and Δy are expressed in pixels. If a shift results in a fixation location outside of the image, the fixation location is repositioned to the nearest possible fixation location. In Fig. 3 'x' represents the current fixation location, and 'o' represents the new fixation location as determined by the neural network.

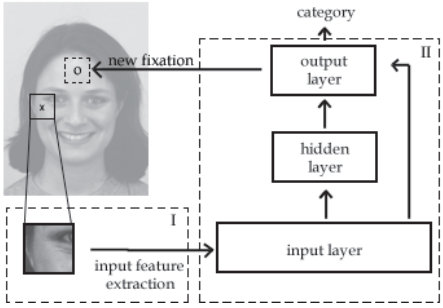

Fig. 3. Overview of the situated model of gaze control.

2.3 Adaptable Parameters

In subsections 2.1 and 2.2 we described the non-situated and the situated model of gaze control. Four types of parameter values define specific instantiations of both models. We refer to these instantiations as agents. The four types of parameters are: the input features, the scale of the raw input window from which features are extracted, the neural network weights, and for the non-situated model the coordinates of all fixation locations. An evolutionary algorithm generates and optimises the agents (i.e., parameter values) by evaluating their performance on the gaze-control task.

3 Experimental Setup

In this section, we describe the gender-recognition task on which we compare the non-situated and the situated model of gaze control (Sect. 3.1). In addition, we discuss the evolutionary algorithm that optimises the models' adaptable parameters (Sect. 3.2). Finally, we mention the experimental settings (Sect. 3.3).

3.1 Gender-Recognition Task

Below, we motivate our choice for the task of gender recognition. Then we describe the data set used for the experiment. Finally, we outline the procedure of training and testing the two types of gaze-control models.

We choose the task of gender recognition in images containing photos of female or male faces, since it is a challenging and well-studied task [9]. There are many differences between male and female faces that can be exploited by gender-recognition algorithms [10, 11]. State-of-the-art algorithms use global features, extracted in a non-situated manner. So far, none of the current algorithms is based on gaze control with a local fixation window.

The data set for the experiment consists of images from J.E. Litton of the Karolinska Institutet in Sweden. It contains 278 images with angry-looking and happy-looking human subjects. These images are converted to gray-scale images and resized to 600×800 pixels.

One half of the image set serves as a training set for both the non-situated and the situated model of gaze control. Both models have to determine whether an image contains a photo of a male or female, based on the input features extracted from the gray-scale images. For the non-situated model, the sequence of T fixation locations is optimised by an evolutionary algorithm. For the situated model, the initial fixation location is defined to be the centre of the image and the subsequent $T-1$ fixation locations are determined by the gaze-shift output values of the neural network (outputs o_2 and o_3). At every fixation, the models have to assign a category to the image. After optimising categorisation on the training set, the remaining half of the image set is used as a test set to determine the performance of the optimised gaze-control models. Both training set and test set consist of 50% males and 50% females.

3.2 Evolutionary Algorithm

As stated in subsection 2.3, an evolutionary algorithm optimises the parameter values that define the non-situated and the situated agents, i.e., instantiations of the non-situated and situated model, respectively. We choose an evolutionary algorithm as our training paradigm, since it allows self-organisation of the closed loop of actions and inputs.

In our experiment, we perform 15 independent 'evolutionary runs' to obtain a reliable estimate of the average performance. Each evolutionary run starts by creating an initial population of M randomly initialised agents. Each agent operates on every image in the training set, and its performance is determined by the following fitness function:

$$f(a) = \frac{t_{c,I}}{IT} \ , \tag{4}$$

in which a represents the agent, $t_{c,I}$ is the number of time steps at which the agent correctly classified images from the training set, I is the number of images in the training set, and T is the total number of time steps (fixations) per image. We note that the product IT is a constant that normalises the performance measure. The $\frac{M}{2}$ agents with the highest performance are selected to form the population of the next generation. Their adaptable parameter sets are mutated with probability P_f for the input feature parameters and P_g for the other parameters, e.g., representing coordinates or network weights. If mutation occurs, a feature parameter is perturbed by adding a random number drawn from the interval $[-p_f, p_f]$. For other types of parameters, this interval is $[-p_g, p_g]$. For every evolutionary run, the selection and reproduction operations are performed for G generations.

3.3 Experimental Settings

In our experiment the models use ten input features. Furthermore, the neural networks of both models have 3 hidden neurons, $h = 3$. All weights of the neural networks are constrained to a fixed interval $[-r, r]$. Since preliminary experiments showed that evolved weights were often close to 0, we have chosen the weight range to be $[-1, 1]$, $r = 1$. The scale of the window from which the input features are extracted ranges from 50 to 150 pixels. Preliminary experiments showed that this range of scales is large enough to allow gender recognition, and small enough for local processing, which requires intelligent gaze control. The situated model's maximal gaze shift m is set to 500, so that the model can reach almost all locations in the image in one time step.

For the evolutionary algorithm we have chosen the following parameter settings: $M = 30$, $G = 300$, and $T = 5$. The choice of T turns out not to be critical to the results with respect to the difference in performance of the two models (see Sect. 4.2). The mutation parameters are: $P_f = 0.02$, $P_g = 0.10$, $p_f = 0.5$, and $p_g = 0.1$.

4 Results

In this section, we show the performances of both models (Sect. 4.1). Then we analyse the best situated agent to gain insight into the mechanism of sensory-motor coordination (Sect. 4.2).

4.1 Performance

Table 1 shows the mean performances on the test set (and standard deviation) of the best agents of the 15 evolutionary runs. Performance is expressed as the proportion of correct categorisations. The table shows that the mean performance of the best situated agents is 0.15 higher than that of the best non-situated agents. Figure 4 shows the histograms of the best performances obtained in the 15 runs for non-situated agents (white) and for situated agents (gray). Since both distributions of the performances are highly skewed, we applied a bootstrap method [12] to test the statistical significance of the results. It revealed that the difference between the mean performances of the two types of agents is significant ($p < 0.05$).

Table 1. Mean performance ($\bar{f}$) and standard deviation (σ) of the performance on the test set of the best agents of the evolutionary runs.

	$\bar{f}$	σ
Non-situated	0.60	0.057
Situated	0.75	0.055

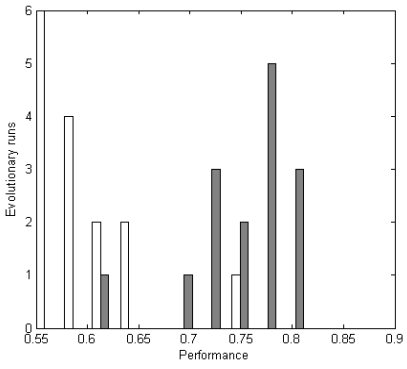

Fig. 4. Histograms of the best fitness of each evolutionary run. White bars are for non-situated agents, gray bars for situated agents.

4.2 Analysis

In this subsection, we analyse the evolved gaze-control policy of the best situated agent of all evolutionary runs. The analysis clarifies how sensory-motor coordination optimises performance on the gender-recognition task.

The first part of the analysis shows that for each category the situated agent controls the gaze in a different way. This evolved behaviour aims at optimising performance by fixating suitable categorisation locations. The second part of the analysis shows that for individual images, too, the situated agent controls the gaze in different ways to fixate suitable categorisation locations.

Gaze Control per Category. Depending on the category, the situated agent fixates different locations. Below, we analyse per category the gaze path of the situated agent when it receives inputs that are typical of that category. The fixations take place at locations that are suitable for categorisation.

To find the suitable categorisation locations per category, we look at the situated agent's categorisation performance on the training set at all positions of a 100×100 grid superimposed on the image. At every position we determine the categorisation ratio for both classes. For the category 'male', the categorisation ratio is: $\frac{c_m(x,y)}{I_m}$, where $c_m(x,y)$ is the number of correctly categorised male images at (x,y), and I_m is the total number of male images in the training set. The left part of Fig. 5 shows a picture of the categorisation ratios represented as intensity for all locations. The highest intensity represents a categorisation ratio of 1. The left part of Fig. 6 shows the categorisation ratios for images containing females. The figures show that dark areas in Fig. 5 tend to have high intensity in Fig. 6 and vice versa. Hence, there is an obvious trade-off between good categorisation of males and good categorisation of females[1]. The presence of a trade-off implies that categorisation of males and females should ideally take place at different locations.

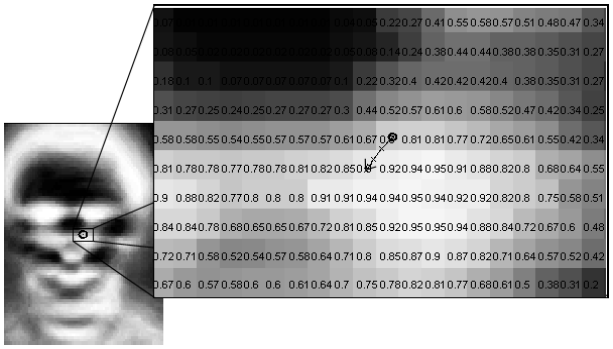

Fig. 5. Categorisation ratios of male images in the training set.

If we zoom into the area in which the agent fixates, we can see that it always moves its fixation location to an area in which it is better at categorising the presumed category. The right part of Fig. 5 zooms in on the categorisation ratios and shows the gaze path that results when the agent receives average male inputs at all fixation locations. The first fixation location is indicated by an 'o'-sign, the last fixation location by an arrow. Intermediate fixations are represented with the 'x'-sign.

[1] Note that the images are not inverted copies: in locations where male and female inputs are very different, good categorisation for both classes can be achieved.

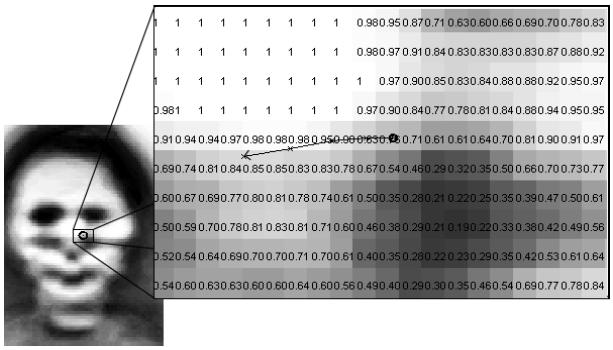

Fig. 6. Categorisation ratios of female images in the training set.

The black lines in Fig. 5 connect the fixation locations. The agent moves from a region with categorisation ratio 0.8 to a region with categorisation ratio 0.9. The right part of Fig. 6 shows the same information for images containing females, revealing a movement from a region with a categorisation ratio of 0.76 through a region with a ratio of 0.98. Both figures show that the situated agent takes misclassifications into account: it avoids areas in which the categorisation ratios for the other category are too low. For example, if we look at the right part of Fig. 5, we see that the agent fixates locations to the bottom left of the starting fixation, while the categorisation ratios are even higher to the bottom right. The reason for this behaviour is that in that area, the categorisation ratios for female images are very low (Fig. 6).

Non-situated agents cannot exploit the trade-off in categorisation ratios. They cannot select fixation locations depending on a presumed category, since the fixation locations are determined in advance for all images.

Gaze Control per Specific Image. The sensory-motor coordination of the situated agents goes further than selecting sensory inputs depending on the presumed category. The categorisation ratios do not explain the complete performance of situated agents. In this section we demonstrate that in specific images, situated agents often deviate from the exemplary gaze paths shown in Fig. 5 and 6 to search for facial properties that enhance categorisation performance.

To see that the categorisation ratios do not explain the complete performance of situated agents, we compare the actual performance of situated agents over time with a predicted performance over time that is based on the categorisation ratios. For the prediction we assume that the categorisation ratios are conditional categorisation probabilities. The categorisation ratio $\frac{c_m(x,y)}{I_m}$ approximates $P_{x,y}(c = M \mid M)$ and $\frac{c_f(x,y)}{I_f}$ approximates $P_{x,y}(c = F \mid F)$. In addition, we assume that conditional probabilities at different locations are independent from each other. We determine the predicted performance of a situated agent by tracking its fixation locations over time for all images and averaging over the conditional categorisation probabilities at those locations. Figure 7 shows both the actual performance (solid lines) as the predicted performance (dotted lines) over time, averaged over all situated agents (squares) and for the best situated agent in particular (circles).

For the last three time steps the actual performances of the situated agents are higher than the predicted performances. The cause of this discrepancy is that the predicted performance is based on the assumption that the conditional categorisation probabilities at different positions are independent from each other. This assumption can be violated, for example, in the case of two adjacent locations.

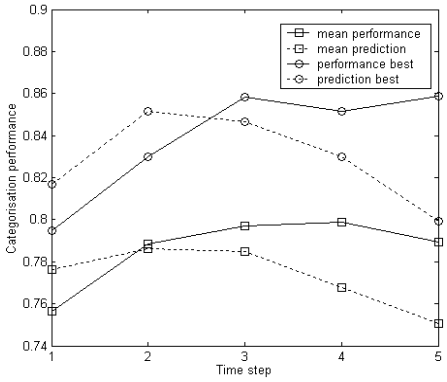

Fig. 7. Actual performance (solid lines) and predicted performance (dotted lines) over time, averaged over all agents (squares) and for the best situated agent in particular (circles).

The situated agent exploits the dependencies by using input features to shift gaze to fixation locations that are well suited for the task of gender recognition. For example, the best situated agent bases its categorisation partly on the eye-brows of a person. If the eye-brows of a male are lifted higher than usual, the agent occasionally fixates a location right and above of the starting fixation. This area is generally not good for male categorisation ($\frac{c_m(x,y)}{I_m} = 0.57$, see Fig. 5), since eye-brows in our training set are usually not lifted. However, for some specific images it *is* a good area, because it contains a (part of a) lifted eye-brow.

The gaze control policy of the situated agents results in the optimisation of (actual) performance over time. Figure 7 shows that the actual performance augments after $t = 1$. The fact that performance generally increases over time reveals that sensory-motor coordination establishes dependencies between *multiple* actions and observations that are exploited to optimise categorisation performance. As mentioned in Sect. 3.3, other settings of T ($T > 1$) lead to similar results. Finally, we remark that for large T ($T > 20$), the performance of the non-situated model deteriorates due to the increased search space of fixation locations.

5 Discussion

We expect our results to generalise to other image classification tasks. Further analysis or empirical verification is necessary to confirm this expectation. Our results may be relevant to two research areas.

First, the results may be relevant to the research area of computer vision. Most research on computer vision focuses on improving pre-processing (i.e., finding appropriate features) and on classification (i.e., mapping the features to an appropriate class) [13]. However, a few studies focus on a situated model (or 'closed-loop model') [14, 15]. Our study extends the application of a situated model using a local input window to the high-level task of gender recognition.

Second, the results are related to research on human gaze control. Of course there is an enormous difference between the sensory-motor apparatus and neural apparatus of the situated model and that of a real human subject. Nonetheless, there might be parallels between the gaze-control policies of the situated model and that of human subjects. There are a few other studies that focus explicitly on the use of situated computational models in gaze control [16, 17], but they also rely on

simplified visual environments. Our model may contribute to a better understanding of gaze control in realistic visual environments.

6 Conclusion

We may draw two conclusions as an answer to the research questions posed in the introduction. First, we conclude that sensory-motor coordination contributes to the performance of situated models on the high-level task of artificial gaze control for gender recognition in natural images. Second, we conclude that the mechanism of sensory-motor coordination optimises categorisation performance by establishing useful dependencies between multiple actions and observations; situated agents search adequate categorisation areas in the image by determining fixation locations that depend on the presumed image category and on specific image properties.

References

1. Pfeifer, R., Scheier, C.: Understanding Intelligence. MIT Press, Cambridge, MA (1999)
2. O'Regan, J.K., Noë, A.: A sensorimotor account of vision and visual consciousness. Behavioral and Brain Sciences **24:5** (2001) 883–917
3. Nolfi, S.: Power and the limits of reactive agents. Neurocomputing **42** (2002) 119–145
4. Nolfi, S., Marocco, D.: Evolving robots able to visually discriminate between objects with different size. International Journal of Robotics and Automation **17:4** (2002) 163–170
5. Beer, R.D.: The dynamics of active categorical perception in an evolved model agent. Adaptive Behavior **11:4** (2003) 209–243
6. van Dartel, M.F., Sprinkhuizen-Kuyper, I.G., Postma, E.O., van den Herik, H.J.: Reactive agents and perceptual ambiguity. Adaptive Behavior (in press)
7. Floreano, D., Kato, T., Marocco, D., Sauser, E.: Coevolution of active vision and feature selection. Biological Cybernetics **90:3** (2004) 218–228
8. Viola, P., Jones, M.J.: Robust real-time object detection. Cambridge Research Laboratory, Technical Report Series (2001)
9. Bruce, V., Young, A.: In the eye of the beholder. Oxford University Press (2000)
10. Moghaddam, B., Yang, M.H.: Learning gender with support faces. IEEE Trans. Pattern Analysis and Machine Intelligence **24:5** (2002) 707–711
11. Calder, A.J., Burton, A.M., Miller, P., Young, A.W., Akamatsu, S.: A principal component analysis of facial expressions. Vision research **41:9** (2001) 1179–1208
12. Cohen., P.: Empirical Methods for Artificial Intelligence. MIT Press, Cambridge, Massachusetts (1995)
13. Forsyth, D.A., Ponce, J.: Computer Vision: a Modern Approach. Prentice Hall, New Jersey (2003)
14. Köppen, M., Nickolay, B.: Design of image exploring agent using genetic programming. In: Proc. IIZUKA'96, Iizuka, Japan (1996) 549–552
15. Peng, J., Bhanu, B.: Closed-loop object recognition using reinforcement learning. IEEE Transactions on Pattern Analysis and Machine Intelligence **20:2** (1998) 139–154
16. Schlesinger, M., Parisi, D.: The agent-based approach: A new direction for computational models of development. Developmental review **21** (2001) 121–146
17. Sprague, N., Ballard, D.: Eye movements for reward maximization. In Thrun, S., Saul, L., Schölkopf, B., eds.: Advances in Neural Information Processing Systems 16. MIT Press, Cambridge, MA (2004)