A Situated Model of Active Vision*

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

IKAT, Universiteit Maastricht, P.O. Box 616, 6200 MD, Maastricht, The Netherlands †

Abstract

There are two main approaches to active vision for classification. The *probabilistic approach* employs a belief state concerning the class of the instance and selects actions to reduce uncertainty in the belief state. The *situated approach* directly learns a mapping from observations to actions (including the classification), driven by the performance on the classification task. One of the advantages of the probabilistic approach is that its action selection takes uncertainty into account. In this paper, we show that the situated approach also takes uncertainty into account in its action selection. To this end, we investigate how a situated gaze-control model without any memory can improve its performance over time in a classification task. We show that it does so by reducing the entropy of the posterior distributions of the possible observations. In this manner, the situated gaze-control model reduces class uncertainty in its observations. In addition, we show that the fixation location of the model serves as a memory of previous observations, which allows the memory-less model to verify conjunctions and disjunctions of facial properties.

1 Introduction

Performance on visual tasks such as classification can be enhanced by employing active vision models. Such models do not passively receive observations, but have to some extent control over the observations they perceive. There are two main approaches to constructing active vision models. In the *probabilistic approach*, active vision models are defined within a probabilistic framework [6]. For example, a classification model starts with uncertainty on the class of the current instance. This uncertainty is modelled explicitly by a belief state representing a probability distribution over all possible classes. The initial belief state is usually a uniform distribution over all possible classes. Subsequently, the model selects actions so that class uncertainty is reduced. Since entropy is a measure of uncertainty, a typical method of action selection is to estimate the entropy of the posterior class distribution after performing an action and to select the action resulting in the minimal entropy [1]. In the *situated approach*, the uncertainty in vision is not explicitly modelled. Instead, the situated approach typically employs an action selection policy that directly maps observations to actions [2]. The action selection policy is optimised for a specific task, for instance by evolutionary algorithms [3],

*This research is carried out within the ToKeN VindIT project (grant number 634.000.018) of the Netherlands Organisation for Scientific Research (NWO).

†g.decroon@cs.unimaas.nl, postma@cs.unimaas.nl, herik@cs.unimaas.nl

only driven by the performance on the task. We refer to this approach as the situated approach, since it relies on finding a policy that exploits the situated nature of the model, i.e., that exploits the closed loop of actions and observations.

Advocates of the probabilistic approach sometimes criticise the situated approach, mainly because of the employment of memory-less policies [6]: *"Obvious limits in perception (e.g., robots cannot see through walls) pose clear boundaries on the type of tasks that can be tackled with this approach."* The limitations of memory-less active vision behaviour have also been discussed by the advocates of the situated approach [4]. An advantage of the probabilistic approach mentioned in [6] is that uncertainty is acknowledged in action selection. For instance, in a classification task, probabilistic active vision models can select actions to reduce class uncertainty and improve classification performance over time.

In [2] it was shown that a memory-less situated model, too, can select actions so that its classification performance increases over time. Apparently, observations are 'remembered' in some way to improve the performance. Since this seems to contradict with the memory-less nature of the model, we here examine the following research question: *How does an active vision classification model of the situated approach increase its performance over time?* To answer the research question we adjust the situated active vision model in [2], so that it is more amenable to analysis, and apply it to the task of gender recognition in natural images. We analyse the model's behaviour within a probabilistic framework to facilitate a comparison with probabilistic active vision models. We explicitly state that we are interested in the manner in which the model improves its performance over time, and not in achieving the best classification performance.

The remainder of the paper is organised as follows. We discuss the active vision model in Section 2. In Section 3 we explain the experimental setup. In Section 4 we analyse the results within a probabilistic framework and draw our conclusion in Section 5.

2 The Model

In Subsection 2.1 we give an overview of the model and we discuss its adaptable parameters in Subsection 2.2.

2.1 Overview

Figure 1 shows an overview of the situated active vision model. It consists of three modules: a sensory module, a controller module, and a classifier module. The first module receives as sensory input the raw input from the window with centre 'x', the current fixation location (see box I in Fig. 1). From that window input features are extracted (see [2] for further details). The vector of real-valued features is mapped onto a limited number of prototypes, with each prototype representing one discrete state. The state is determined with a nearest-neighbour approach, i.e., the prototype with the smallest distance to the extracted feature vector represents the state. The prototypes are optimised by the evolutionary algorithm (see Subsection 2.2). The second module is a controller that maps the sensory state (as calculated by the first module) to an action, i.e., a gaze shift in

the image. The controller is a state-action table that assigns a value in the range [-1, 1] to each state-action pair. It executes the action that has the highest value in the current state. All gaze shifts are restricted to a grid around the current fixation location, as illustrated in Figure 2. 'w' and 'h' indicate the maximal horizontal and vertical displacement. The arrow illustrates a possible gaze shift. At the new fixation location, new sensory inputs are received by the first module. The third module is a classifier that maps the sensory state to a class. The classifier is a state-class table that assigns a value in the range [-1, 1] to each state-class pair. The classifier only classifies the last sensory state of a fixation sequence, by selecting the class with the highest value for that state.

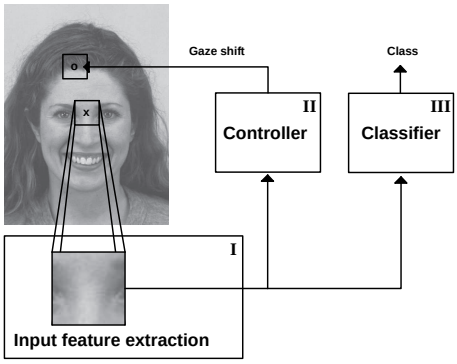

Figure 1: Overview of the active vision model. An 'x' marks the current fixation location. The three modules that constitute the model are illustrated by the boxes 'I', 'II', and 'III'.

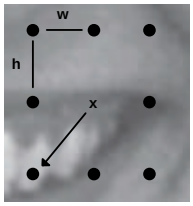

Figure 2: Grid indicating the locations to which the model can shift its gaze from the current fixation location 'x'.

2.2 Adaptable Parameters

Five parameter types will define specific instantiations of the model. We refer to these instantiations as agents. The five parameter types are: the input features, the scale of the raw input window from which features are extracted, the prototypes, the values in the controller's state-action table, and the values in the classifier's state-class table. An evolutionary algorithm generates and optimises the agents (i.e., parameter values) by evaluating their performance on the recognition task.

3 Experimental setup

In this section, we describe the gender-recognition task to which we apply the active vision model (Subsection 3.1). In addition, we discuss the evolutionary algorithm that optimises the model's adaptable parameters (Subsection 3.2). Finally, we indicate the experimental settings (Subsection 3.3).

3.1 Gender Recognition Task

We apply the active vision model to the gender-recognition task described in [2]. The data set for the experiment consists of images from J.E. Litton of the Karolin-

ska Institutet in Sweden. It contains 278 images with angry-looking and happy-looking human subjects. The 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 the situated model of gaze control, that has to determine whether an image contains a male or a female face, based on the input features extracted from the gray-scale images. The initial fixation location of the active vision model is defined to be the centre of the image. The subsequent $(T-1) \geq 1$ fixation locations are determined by the controller. At the last fixation the model assigns a class to the image. After optimising classification 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 model. Both training set and test set consist of 50% males and 50% females.

3.2 Evolutionary Algorithm

An evolutionary algorithm optimises the parameter values that define the situated agents, i.e., instantiations of the situated model. 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: $f(a) = c/I$, in which a represents the agent, c is the amount of correctly classified images in the training set, and I is the total number of images in the training set. The $M/2$ 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 prototypes or state-action table values. If mutation occurs, a feature parameter is perturbed by adding a random number drawn from the interval $[-p_f; p_f]$. For the 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 model uses ten input features. Furthermore, the model has a list of ten prototype feature vectors, corresponding to ten distinct states. Preliminary experiments indicated this number of vectors to be sufficient. The scale of the window from which the input features are extracted ranges from 50 to 150 pixels. Furthermore, $h = 38$ and $w = 28$ pixels, so that there are $20 \times 20 = 400$ grid points in an image that can be fixated. The evolutionary algorithm has the following parameter settings: $M = 30$, $G = 300$, and $T = 4$. The mutation parameters are: $P_f = 0.02$, $P_g = 0.10$, $p_f = 0.5$, and $p_g = 0.1$.

4 Results

We first discuss the performance of the active vision model in Subsection 4.1 and then analyse its behaviour in Subsection 4.2.

4.1 Performance

The results of the agents are expressed in the proportion of correctly classified images on the test set. The performance of the active vision model averaged over 15 independent evolutionary runs was 74,88%, with a standard deviation of 5,75%. As expected, the performance of all agents increases over time, implying that the adjusted active vision model exhibits sensible gaze behaviour.

4.2 Analysis

In this subsection, we analyse a typical evolved agent to study how it achieves an increase in performance over time. As described in Section 3, the evolutionary algorithm optimises the prototype feature vectors that determine the sensory states. During execution of the task, the agent only encounters three of the ten possible sensory states (henceforth referred to as o_1, o_2, and o_3). Apparently, seven out of ten prototype feature vectors are never close to the extracted feature vector. The classifier maps o_1 to 'female', and both o_2 and o_3 to 'male'. The controller shifts the gaze to the left in the case of o_1 and o_2, and to the top left in the case of o_3.

4.2.1 Probabilistic Behavioural Explanation

We first describe the general behaviour of the agent, and then we analyse the agent's behaviour within a probabilistic framework.

The first fixation location at the centre of the image is usually located at the eye-brows. Darker and thicker eye-brows result in o_1 or o_2, so that the agent shifts its gaze to the left. For such images, the gaze keeps shifting to the left for the remaining three time steps, generally ending with o_2 (associated with male images) at the final time step. If the eye-brows are lighter, thinner, or not present at the first fixation location, the resulting observation is o_3. Consequently, the gaze shifts to the top left, located on the forehead. This leads to a new observation of o_3. Only if the agent fixates the hair(line) of a person, the observation changes to o_1, the observation associated with female images. The controller maps this observation to a gaze shift to the left, into to the hair of the person, which leads once more to an observation of o_1. If the agent does not reach the hairline within 4 time steps, the observation is still o_3, associated with male images. In other words, for male images the agent verifies a disjunction of facial properties (dark eye-brows or a high hairline) and for female images a conjunction of facial properties (light eye-brows and a low hairline). Since the conditions of the conjunction or disjunction are verified at different time steps, the performance increases over time.

In terms of a probabilistic framework, the controller and the classifier have two different roles. The role of the classifier is to act as a Maximum A Posteriori classifier (MAP-classifier) at the last time step. A MAP-classifier is a classifier that maps an observation to the most probable class. The performance of a MAP-classifier is negatively correlated to the entropy of the posterior distribution. In our experiments, the entropy is maximal if an observation can equally probably be caused by a male as by a female image. It is minimal if an observation can only be caused by either a male or a female image. In fact, the role of the agent's controller is to minimise the entropy of the posterior probability distribution over

time. We show this by calculating the entropy for every time step. We define the entropy E at a time step as the weighted sum of the Shannon entropies [5] for all different observations at that time step, as follows:

$$E = \sum_{o} P(o)\ H(P(C \mid o)) \tag{1}$$

$$H(P(C \mid o)) = \sum_{c} P(c \mid o) \log_2(\frac{1}{P(c \mid o)}) \tag{2}$$

in which $P(o)$ is the probability for observation o and $H(P(C \mid o))$ is the Shannon entropy of the posterior probability distribution for observation o. Furthermore, C is the set of all mutually exclusive classes c. Figure 3 shows the entropy over time of the analysed agent (dotted line). It slightly increases from $t = 1$ to $t = 3$, and then decreases from $t = 3$ to $t = 4$. Note that this policy can only be found with non-greedy search: greedy search would never result in an increase of entropy. Figure 3 also shows the performance of the agent (solid line) and the performance of a MAP-classifier (dashed line) over time. The MAP-classifier's performance first decreases from $t = 1$ to $t = 3$, and then increases from $t = 3$ to $t = 4$. It outperforms the agent's classifier on the first three time steps, but has a performance equal to the agent's classifier at the last time step. Hence, the agent's classifier is a MAP-classifier for the last time step. Figure 4 shows the same information as Figure 3, but then averaged over all agents. It shows that the findings for the analysed agent are representative for all evolved agents, although the entropy generally decreases more smoothly. These results show that situated agents reduce the entropy of the posterior probability distribution, i.e., the class uncertainty in the observations, over time, without employing belief states.

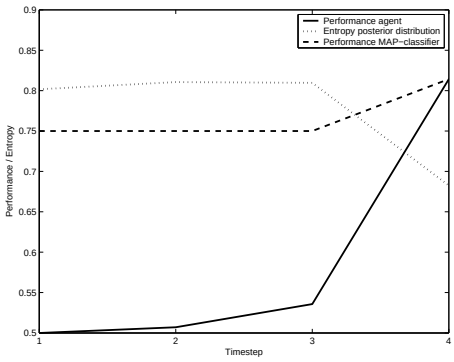

Figure 3: Entropy, agent performance, and MAP-classifier performance over time of the analysed agent on the training set.

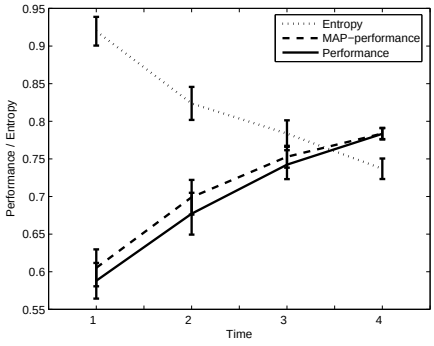

Figure 4: Entropy, agent performance, and MAP-classifier performance over time on the training set, averaged over all evolutionary runs.

4.2.2 Memory

The analysis above indicates that observations are 'remembered' in some way to improve performance over time. This is surprising, since the controller and classifier are both tables that perform a memory-less mapping from sensory states to actions or classifications, respectively. In this section, we show that the agent exploits an 'external memory' (cf. [7]).

Figure 5 shows the fixation locations of the agent on the whole training set, with all different gaze paths exhibited by the agent. The background image is included to give a rough impression of how the fixation locations relate to facial features. It is the mean image of four training set images, two male and two female images. The figure illustrates that the fixation location contains information on the observation history. For instance, the fixation location indicated with a '*' is only reached if the observation history for the first three time steps is (o_3, o_3, o_3), observations that result in three gaze shifts to the top left. Therefore, the fixation location can be regarded as a memory of the preceding observations. Since this memory is not part of the agent's internal mechanism, but part of the state of the world, it can be regarded as an external memory.

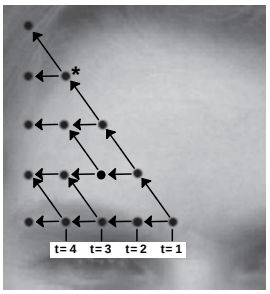

Figure 5: Fixation locations of the agent on the whole training set, with all different gaze paths.

The observation history influences the observation at the last time step, since the fixation location determines the probability distribution of the observations. In other words, each observation depends on the preceding observations by means of the external memory. To show that there is such a dependence, it is sufficient to analyse the first two time steps. At the first time step, the probabilities (as estimated with the training set) for the different observations perceived by the agent are: $P(o_1(1)) = 0.04$, $P(o_2(1)) = 0.47$, and $P(o_3(1)) = 0.49$. The probabilities at the second time step are $P(o_1(2)) = 0.02$, $P(o_2(2)) = 0.51$, and $P(o_3(2)) = 0.47$. Hence, the probabilities for the observations at the first two time steps are rather similar. However, observations at different time steps depend on each other by means of the fixation locations. For example, given an observation o_2 at the second time step, it is very probable that the agent also observed o_2 at the first time step: $P(o_2(1) \mid o_2(2)) = 0.92$, $P(o_1(1) \mid o_2(2)) = 0.08$, and $P(o_3(1) \mid o_2(2)) = 0$. Since $P(o_2(1) \mid o_2(2)) > P(o_2(1))$, the observation at the second time step is a form of memory. In the same manner, $o_1(1)$ and $o_3(1)$ are remembered in the

observation at $t = 2$; $P(o_3(1) \mid o_1(2)) = 1$, and $P(o_3(1) \mid o_3(2)) = 1$. The cause of this high dependence between the second and the first observation, is that the observation probability distributions are very different at the two fixation locations for $t = 2$, and that the agent fixates one of those locations on the basis of the first observation. In a similar fashion, the observations at the last time step are dependent on the whole sequence of previous observations.

5 Conclusion

From the experiments, we may conclude that the situated active vision model increases its performance over time by reducing the entropy of the posterior distributions of the possible observations with its controller. This implies that the situated gaze-control model reduces class uncertainty in its observations, instead of reducing class uncertainty in a belief state as probabilistic models. The fixation location of the model serves as an external memory of previous observations, which allows the model to verify conjunctions and disjunctions of facial properties.

We envisage two directions of future research. First, we want to compare a model of the probabilistic approach more closely with a model of the situated approach. One interesting aspect about such a comparison is that most probabilistic models of active vision assume independence of subsequent observations, while the situated model discussed in this paper bases its policy on the dependencies between observations (see the analysis in Section 4). Second, we want to study situated models of gaze control that have an internal memory. A possible research question on such models is, whether they develop a mechanism that acts similarly to a belief state.

References

[1] T. Arbel and F.P. Ferrie. Entropy-based gaze planning. *Image and Vision Computing*, 19(11):779 – 786, 2001.

[2] G. de Croon, E.O. Postma, and H.J. van den Herik. Sensory-motor coordination in gaze control. In F. Rothlauf, editor, *Applications of Evolutionary Computing - EvoWorkshops 2005*, Lausanne, Switzerland, 2005.

[3] J. Holland. Genetic algorithms. *Scientific American*, pages 66–72, 1992.

[4] S. Nolfi. Power and the limits of reactive agents. *Neurocomputing*, 42:119–145, 2002.

[5] C.E. Shannon. A mathematical theory of communication. *The Bell System Technical Journal*, 27:379–423, 623–656, 1948.

[6] S. Thrun. Probabilistic algorithms in robotics. *AI Magazine*, 21(4):93–109, 2000.

[7] M. F. van Dartel, I. G. Sprinkhuizen-Kuyper, E. O. Postma, and H. J. van den Herik. Reactive agents and perceptual ambiguity. *Adaptive Behavior*, in press.