Comparing Active Vision Models

G. de Croon   I.G. Sprinkhuizen-Kuyper   E.O. Postma

July 17, 2006

Abstract

Performance on visual tasks such as classification can be enhanced by employing active vision systems. Such systems do not passively receive observations, but have to some extent control over the observations they perceive. There are two general approaches to active vision. The first approach to active vision is a probabilistic approach, in which reducing uncertainty on a part of the world state is the central goal. This uncertainty is modelled by a belief state. The second approach to active vision is a behavioural approach, in which successful behaviour is the central goal. For both approaches, there have been considerable research efforts into designing and studying various active vision models. However, it is not clear how the different existing active vision models relate to each other, and what their relative advantages are. In this report, we identify three main types of active vision models in the probabilistic approach and describe them in a common formal framework. The first type of model selects actions on the basis of the mutual information between actions and classes, and is referred to as the Mutual Information model (MI). The second type of model learns an action policy on the basis of entropy loss in the belief state, the Entropy Loss model (EL). The third type of model bases its action selection on the mode of the belief state, the Mode of Belief state model (MB). In addition, we introduce a fourth type of active vision model that is based on the behavioural approach to active vision, the BeHavioural model (BH). Model BH is identical to EL, except that it learns an action policy that achieves a high performance rather than one that achieves entropy loss in the belief state. We compare the four active vision models empirically on a view-based three-dimensional object classification task. The experimental results give insight into the differences between the models. The overall result is that BH generally outperforms the models EL and MB of the probabilistic approach. In addition, within the probabilistic approach model MI has the best classification performance. Besides revealing performance differences between the active vision models, the experimental results also illustrate properties of the relation between the usefulness of active vision, the number of objects involved in the classification task, and the richness of the visual observations of the models.

1 Introduction

Performance on visual tasks such as classification can be enhanced by employing active vision systems. Such systems do not passively receive observations, but have to some extent control over the observations they perceive.

There are two general approaches to active vision. The first approach to active vision is a probabilistic approach, in which reducing uncertainty on a part of the world state is the central goal (e.g., [10]). For example, the model might have to determine the true class of a certain object, where it starts with uncertainty on the class of the current instance. This uncertainty is modelled by a belief state. A belief state is a probability distribution over all possible classes. The model selects actions so that the resulting observations allow a reduction of class uncertainty. Since entropy is a measure of uncertainty, a typical method of action selection is to select the action resulting in the minimal entropy of the belief state (e.g., [1]). In the probabilistic approach to active vision, we discern different types of active vision models that differ in their action selection strategies and sometimes also in their belief state updates. Although much research has been devoted to designing and studying these types of active vision models [1, 3, 10, 16], so far most of the models have only been compared with a random action strategy.

The second approach to active vision is a behavioural approach, in which successful behaviour is the central goal (e.g., [6, 12]). This approach can be applied to a more general type of problem. For example, it can be applied to behavioural classification tasks in which the model has no belief state regarding the class, or even no memory of past observations [7, 12]. An example of a behavioural task is the task in [7], in which an active vision model has to catch small blocks and avoid large blocks. The behavioural approach does not necessarily strive for reducing entropy maximally. The only goal is to select actions and observations as to maximise the performance. Behavioural models of active vision have not yet been compared to probabilistic models of active vision.

It is unclear how different types of active vision models relate to each other, and what the advantages of those models are with respect to each other. In this report, we want to clarify this matter by comparing different active vision models. We identify three main types of probabilistic active vision models and describe them in a common formal framework. In addition, we introduce a fourth type of active vision model that is based on the behavioural approach to active vision. Subsequently, we compare the four active vision models empirically on a classification task of 3-D objects. In our comparison of the models, we focus on the differences in the action selection strategies of the four active vision models. Therefore, we provide all active vision models with the same belief state update.

The remainder of this report is organised as follows. In Section 2, we introduce the notation used throughout the report. Then, in Section 3 we explain the belief state update that is used by all active vision models described in this report. We describe the active vision models in Section 4. In Section 5 we discuss the experimental setup with which we compare the models. Then, in

Section 6 we show the results of these experiments. In Section 7 we perform an analysis that provides an explanation for some of the experimental results. We discuss our experiments and draw our conclusions in Section 8.

2 Notation

Throughout the report we will use the following notation. With $p(X)$ we indicate a probability distribution over all elements of the set X. Where capital letters represent sets, small letters represent elements. Therefore $p(x)$ is the probability of a specific element $x \in X$. Since we discuss all active vision models in the context of a classification task, we use the following variables: O is the set of all possible observations, C the set of all possible classes, and A the set of all possible actions. A specific observation, class, and action are denoted by o, c, and a respectively, where $o \in O$, $c \in C$, and $a \in A$. In our discussion of the models, we will assume the variables O, C, and A to be discrete, finite sets, but all models can be applied to continuous variables as well (see, e.g., [10]). The active vision process extends itself over multiple discrete time steps. The time step before the model takes an action, or performs an observation, is indicated with 0. A possible maximum number of time steps is indicated with t. The current time step is represented by i. A bold letter with subscript j indicates a sequence of length j. In this report typically the subscript i is used, where for example $\mathbf{o}_i = \langle o_1, o_2, \ldots, o_i \rangle$ is the sequence of observations of the first i time steps. Similarly, the sequence of corresponding actions is $\mathbf{a}_i = \langle a_1, a_2, \ldots, a_i \rangle$.

The belief state at a time step i represents the probabilities of the classes, given the past observations and actions: $p(C \mid \mathbf{o}_{i-1}, \mathbf{a}_{i-1})$. The probability of a particular class c at time step i is indicated with $p(c \mid \mathbf{o}_{i-1}, \mathbf{a}_{i-1})$. On the basis of the current belief state $p(C \mid \mathbf{o}_{i-1}, \mathbf{a}_{i-1})$, the probabilistic active vision models determine an action for the current time step, a_i, which results in an observation o_i. The action and observation allow the models to calculate the belief state for the next time step, $p(C \mid \mathbf{o}_i, \mathbf{a}_i)$. At time step $i = 1$, the belief state is equal to the prior distribution, i.e., $p(C \mid \mathbf{o}_0, \mathbf{a}_0) = p(C)$, since $\mathbf{o}_0 = \mathbf{a}_0 = \langle\rangle$.

3 Belief State Update

Since we want to focus on the differences in action selection strategies between the different types of active vision models, we provide all models with the same belief state update rule. In this section, we explain how all active vision models update their belief state in our experiments. We place our explanation of the belief state update in the context of a classification task.

The recursive belief state update presented here, was introduced in [9, 10]. We select this particular belief state update, since it can be employed by all four active vision models that we introduce in Section 4. In the following, we derive the recursive update for the belief state, as used in [10]. The belief state is the posterior class probability distribution, which we can rewrite using Bayes' rule as follows.

$$p(C \mid \mathbf{o}_i, \mathbf{a}_i) = \frac{p(o_i \mid C, \mathbf{o}_{i-1}, \mathbf{a}_i) p(C \mid \mathbf{o}_{i-1}, \mathbf{a}_i)}{p(o_i \mid \mathbf{o}_{i-1}, \mathbf{a}_i)} \tag{1}$$

In [10] the classification task concerns the classification of different 3-D objects. Each action of the active vision model corresponds to an angle from which an object can be viewed. Since in their experimental setup any angle can be reached at any time step, it is assumed that an observation is only determined by the class and the action, and not by the past observations and actions:

$$p(o_i \mid C, \mathbf{o}_{i-1}, \mathbf{a}_i) = p(o_i \mid C, a_i) \tag{2}$$

This assumption can be used to rewrite the formula of the posterior class probability distribution as follows.

$$p(C \mid \mathbf{o}_i, \mathbf{a}_i) = \frac{p(o_i \mid C, a_i) p(C \mid \mathbf{o}_{i-1}, \mathbf{a}_i)}{p(o_i \mid \mathbf{o}_{i-1}, \mathbf{a}_i)} \tag{3}$$

The formula can further be simplified by using the fact that $p(C \mid \mathbf{o}_{i-1}, \mathbf{a}_i) = p(C \mid \mathbf{o}_{i-1}, \mathbf{a}_{i-1})$, since action a_i does not contain any information on the class if o_i is unknown. In consequence, the denominator $p(o_i \mid \mathbf{o}_{i-1}, \mathbf{a}_i)$ can be rewritten as follows.

$$p(o_i \mid \mathbf{o}_{i-1}, \mathbf{a}_i) = \tag{4}$$

$$\sum_{c \in C} p(c \mid \mathbf{o}_{i-1}, \mathbf{a}_i) p(o_i \mid c, \mathbf{o}_{i-1}, \mathbf{a}_i) =$$

$$\sum_{c \in C} p(c \mid \mathbf{o}_{i-1}, \mathbf{a}_{i-1}) p(o_i \mid c, \mathbf{o}_{i-1}, \mathbf{a}_i) =$$

$$\sum_{c \in C} p(c \mid \mathbf{o}_{i-1}, \mathbf{a}_{i-1}) p(o_i \mid c, a_i) \tag{5}$$

The rewritten formula of the posterior class probability distribution becomes:

$$p(C \mid \mathbf{o}_i, \mathbf{a}_i) = \frac{p(o_i \mid C, a_i) p(C \mid \mathbf{o}_{i-1}, \mathbf{a}_{i-1})}{\sum_{c \in C} p(c \mid \mathbf{o}_{i-1}, \mathbf{a}_{i-1}) p(o_i \mid c, a_i)} \tag{6}$$

This formula can be used as a recursive belief state update. Namely, to calculate Equation 6, we only need the old belief state and the observation probability distributions for all classes and actions. The observation probability distributions $p(O \mid c, a)$ for all classes and actions have to be determined before application of the active vision model. The subscript i is dropped here, since it is assumed that these observation probability distributions are static. In this report, all active vision models employ the belief state update of Equation 6.

4 Active Vision Models

In this section, we introduce the four active vision models that we compare in this report. The first three active vision models are based on the probabilistic approach to active vision. As mentioned in the introduction, there are several different active vision models in the probabilistic approach [1, 9, 10, 14, 16]. However, we can discern three main model types, according to how they select actions. The first type of model calculates at every time step the expected usefulness of all actions to select one of them (see, e.g., [3, 4, 9, 10, 11]). This type of model determines the expected usefulness of an action by generating observations according to the observation probability distribution that is associated with that action. These generated observations are used to perform tentative belief state updates. The model selects the action whose tentative belief state updates give the best result. In this report, we implement the active vision model of the first type on the basis of [9, 10], in which the action is selected to maximise the mutual information between observations and classes. For this reason, we refer to this model as 'MI' (Section 4.1). This type of model calculates the mutual information of all possible actions, before selecting one of them. On the contrary, the second type of model (see, e.g., [8, 16]) learns an action selection policy. This policy maps the current belief state to an action. The goal in learning the policy is entropy loss in the belief state. For this reason, we refer to this method as 'EL' (Section 4.2). The third type of model ranks all actions on forehand. These actions are ranked according to the uncertainty reduction in the belief state, assuming that the object belongs to the most probable class (see e.g., [1, 14, 19]). Since the most probable class corresponds to the mode of the belief state, we refer to this model as 'MB' (Section 4.3). We also incorporate a behavioural model in our comparison. As mentioned in the introduction, behavioural models can have an implicit belief state, or no belief state at all. However, in our comparison we want to focus on the main aspect of behavioural models: that their actions are directed towards successful behaviour. Therefore, we introduce a fourth active vision model that does have an explicit belief state, but that takes actions as to achieve a high classification performance. In particular, the fourth model is equal to EL, except that this model's learning goal is to maximise classification performance instead of reducing entropy in the belief state. Since it is the only model that is based on the behavioural approach, we refer to this fourth type of model as 'BH' (Section 4.4). In the rest of this section, we first present the above-described four active vision models. Then, in Section 4.5, we introduce two benchmark models: an active vision model that takes random actions, 'RA', and a passive vision model, 'PA'.

4.1 Action Selection Based on Mutual Information (MI)

Several probabilistic active vision models evaluate all possible actions, before selecting one action for actual execution [3, 4, 9, 10, 11]. The first type of model we discuss, introduced in [10], calculates the mutual information I between observations O and classes C for each action a:

$$I(C;O \mid a) = H(C \mid a) - H(C \mid O,a) = H(O \mid a) - H(O \mid C,a) \tag{7}$$

$$= H(O \mid a) - \sum_{c \in C} p(c \mid a) H(O \mid c,a)$$

$$= -\sum_{o \in O} p(o \mid a) \mathrm{log}(p(o \mid a)) + \sum_{c \in C} p(c \mid a) \sum_{o \in O} p(o \mid c,a) \mathrm{log}(p(o \mid c,a))$$

$$= -\sum_{c \in C} \sum_{o \in O} p(o,c \mid a) \mathrm{log}(p(o \mid a)) + \sum_{c \in C} \sum_{o \in O} p(o,c \mid a) \mathrm{log}(p(o \mid c,a))$$

$$= \sum_{c \in C} \sum_{o \in O} p(o,c \mid a) \mathrm{log}(\frac{p(o \mid c,a)}{p(o \mid a)})$$

$$= \sum_{c \in C} \sum_{o \in O} p(c \mid a) p(o \mid c,a) \mathrm{log}(\frac{p(o \mid c,a)}{p(o \mid a)}) \tag{8}$$

In the context of our classification task, we want to maximise the mutual information between O and C, by selecting an action a_i:

$$a_i = \mathrm{argmax}_a I(C;O \mid \mathbf{o}_{i-1}, \mathbf{a}_{i-1}, a) \tag{9}$$

Where, according to Equation 8:

$$I(C;O \mid \mathbf{o}_{i-1}, \mathbf{a}_{i-1}, a) = \tag{10}$$

$$\sum_{c \in C} \sum_{o \in O} p(c \mid \mathbf{o}_{i-1}, \mathbf{a}_{i-1}, a) p(o \mid c,a) \mathrm{log}(\frac{p(o \mid c,a)}{\sum_{c \in C} p(c \mid \mathbf{o}_{i-1}, \mathbf{a}_{i-1}) p(o \mid c,a)})$$

Note that $p(c \mid \mathbf{o}_{i-1}, \mathbf{a}_{i-1}, a) = p(c \mid \mathbf{o}_{i-1}, \mathbf{a}_{i-1})$, since the probabilities of C are independent of the action that has not yet been executed. Equation 10 shows that the mutual information for an action can be calculated on the basis of the belief state and the observation probability distributions. This active vision model does not need any training time other than the time to learn the observation probability distributions $p(O \mid c,a)$. The execution time of the model depends on both the number of objects and the number of actions. As a consequence, the model takes an increasing amount of execution time for increasing object subset sizes.

4.2 Learning an Action Policy for Entropy Loss (EL)

The second type of active vision model, EL, learns an action selection policy (e.g., [16]). For example, a policy Π can be learned that maps the belief state to an action, $\Pi : p(C \mid \mathbf{o}_i, \mathbf{a}_i) \rightarrow A$. The mapping is learned with reinforcement learning [20], with the entropy reduction in the belief state as the reinforcement signal. In this report we optimise the mapping Π with an evolutionary algorithm [2] instead of with reinforcement learning. We implement Π with a neural network, whose weights form the genome of the policy. The fitness *fit* of policies in the population during evolution is defined as the expected total entropy reduction in the belief state over all t time steps:

$$\mathit{fit}(\Pi) = E[\sum_{i=1}^{t} (H(C \mid \mathbf{o}_{i-1}, \mathbf{a}_{i-1}) - H(C \mid \mathbf{o}_i, \mathbf{a}_i))] = E[H(C) - H(C \mid \mathbf{o}_t, \mathbf{a}_t)] \tag{11}$$

We approximate the expected total entropy reduction by executing a policy Π on all objects in the selected object set and averaging over the measured total entropy reduction. A key property of EL is that its action policy allows very quick action selection at execution time, because it allows the immediate selection of one action without evaluating all actions. However, this quick execution time comes at the cost of a long training time. Also note that where MI always selects the action that results in the largest immediate reduction of entropy, this is not necessarily the case for EL. Therefore, EL has the ability to find non-greedy action selection strategies.

4.3 Action Selection for Most Probable Class (MB)

The third type of active vision model, MB, bases its action selection on the most probable class, i.e., the mode of the belief state. An example of this strategy is [1], in which for every class and every action, the expected entropy of the belief state is estimated before execution. The actions $a \in A$ are ranked for each class $k \in C$ according to the expected entropy reduction, given a uniform belief state:

$$\text{value}(a_k) = E[H(C) - H(C \mid a) \mid k] \tag{12}$$

$$= H(C) - E[H(C \mid a) \mid k] \tag{13}$$

We calculate the expected entropy in Equation 13 by sampling repeatedly an observation $o \in O$ from class k, according to $p(O \mid k, a)$, resulting in observations $o_i(a, k)$, $i = 1, 2, \ldots, m$. Then we use the following approximation:

$$E[H(C \mid a) \mid k] \approx \frac{1}{m} \sum_{i=1}^{m} H(C \mid a, o_i(a, k)) \tag{14}$$

During execution, the model receives an observation, updates the belief state, and selects the best action for the most probable class. It selects the action with the highest value that has not been performed before. Compared to EL, MB has a short training time, since it calculates only once for every action-class pair how good the action is. Compared to MI it has a short execution time, since the ranking of actions resulting from the training procedure allows a fast execution time. However, the advantages concerning training and execution time may come at the cost of a lower performance attained on the task.

4.4 Learning an Action Policy for Performance (BH)

The fourth active vision model, BH, is inspired on the behavioural approach to active vision. The model is equal to the second active vision model that learns an action policy Π that maps the belief state to an action, $\Pi : p(C \mid \mathbf{o}_i, \mathbf{a}_i) \to A$. The only difference with EL is that we use a different fitness function for this model:

$$fit(\Pi) = \frac{k}{n} \tag{15}$$

Where k is the number of correctly classified objects, in a trial in which the model had to classify n objects ($n \geq k$). The reason that we state that this model belongs to the behavioural approach to active vision, is that the central goal of the model's action selection is performance instead of entropy reduction. Again, we implement the action policy Π with a feedforward neural network. As EL, it has a long training time and a short execution time.

4.5 Benchmark Models

In each experiment, we compare the active vision models with two additional models that provide benchmark performances for the experiment. The first additional model is an active vision model that selects actions at random. We refer to this model as 'RA'. It provides a bottom performance for the active vision models.

The second additional model represents a passive vision model. This is a model that takes one random action and immediately classifies the object with the help of the observation probability distribution. It selects the class that has the highest probability to generate the observation o, given its randomly selected a. We refer to this model as 'PA', which stands for passive.

5 Setup Empirical Comparison

Most of the probabilistic active vision models have been introduced in the context of a classification task of 3-D objects [1, 3, 10, 16]. Therefore, for our empirical comparison of the active vision models, we also employ such a task. For the experiments, we use the ALOI-data set [13] that consists of 1000 objects that have been placed on a turn-table and have been photographed from 72 angles that are each 5° apart. The resulting images are gray-value images. We use the set of objects with the smallest image size (192 × 144) for our experiments. To ensure a fair comparison between the approaches, we divide the data set into two subsets: a subset of 250 objects that is used to tune the parameters of the various active vision models, and a subset of 750 objects that is used to do the experimental comparison of all (tuned) active vision models.

Here we give an overview of the procedure of a single experimental run, that we explain in more detail below. First, a subset of n objects is selected at random from the current object set (tuning or comparison set) and the corresponding images are loaded and resized to one fourth of their size with bicubic resampling. Figure 1 shows an example subset of 25 objects. Following the experimental setup in [18], we divide the images in a training set (images with angles divisible by 5°, excluding those divisible by 10°) and a test set (remaining images). As in [3, 10, 16], we apply Principal Component Analysis to the training images of the n selected objects (Subsection 5.1). We use the extracted principal components to estimate the probability distributions $p(O \mid a, c)$ for all actions and classes (Subsection 5.2). For the models EL, MB, and BH we use these probabilities for training the action selection (Subsection 5.3). As noted in Section 4.1, MI does not require training. For the action selection, we follow the setup in [10], in which an action corresponds to selecting one of the angles from which the object can be viewed. Finally, we apply the trained model to all object-images in the test set, and measure its performance (Subsection 5.4).

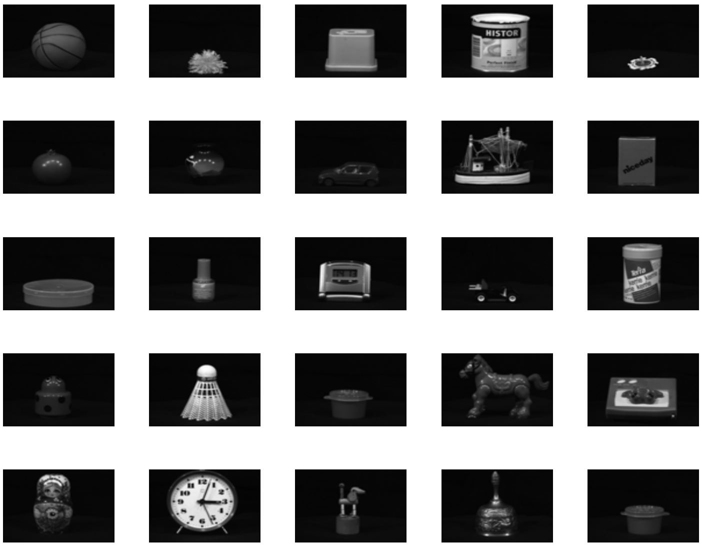

Figure 1: Example set of objects for $n = 25$.

When an active vision model has to classify an object, it starts with a uniform belief state. We perform two types of experiments. In the first type, the active vision model can determine the first angle itself. In the second type, the run starts with an observation from a random angle, unknown to the active vision model. In both types of experiments, the belief state is updated on the basis of the first observation. As in [10], the active vision model continues selecting angles until the confidence in one of the classes is higher than 90% ($max(p(c)) \geq 0.90$) or the number of observations is equal to 10. If the number of observations is equal to 10, the class k with highest probability is selected: $k = argmax_c p(c \mid \mathbf{o}_i, \mathbf{a}_i)$.

5.1 Observations

The observations for the classification task are based on Principal Component Analysis (PCA, see e.g., [15]). We apply PCA to the training images in order to map them to a lower-dimensional space that retains much of the original data's variance. Before applying PCA we transform each image to a vector by concatenating the image's rows of gray values. For finding the principal components, we implemented the iterative version of simple PCA [17], since it is fast and well-suited for large image sets. Figure 2 shows a visualisation of the first five principal components of the object subset shown in Figure 1. To obtain these visualisations, we have applied simple PCA to the training images of all objects in the subset. Then, we scaled the values of the eigenvectors so that all its values are in the range $[0, 1]$. In the visualisation of Figure 2, we have retransformed the vectors to two-dimensional matrices and the rescaled values represent light intensity, where 0 is black and 1 is white. The visualisation of the first principal component shows that most variance of light intensity in the example object subset is located in the middle of the image, i.e., the point around which the objects turn. The visualisation of the second and third principal components show a variation in the top and bottom of the objects. In the experiments, after finding the principal components, we obtain an observation o of an image by transforming the image to a vector and projecting it onto the principal components.

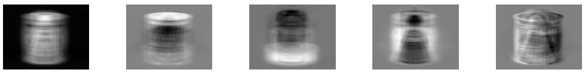

Figure 2: Visualisation of the first five principal components of the object set shown in Figure 1.

5.2 Observation Probability Distributions

In order to update the belief state, we need to estimate the probabilities $p(O|A, C)$. For simplicity, we assume, as in [3, 10, 14], that the observations O for a given action a and class c are Gaussian distributed in the space spanned by the principal components: $p(O|a,c) = N(\mu, \Sigma)$, with μ and Σ dependent on a and c. The covariance matrix Σ is assumed to be diagonal. The parameters of the distribution are estimated for each combination of a and c with the help of the training set. Since only one image is available per class and angle, we perturb the images in order to obtain the observation probability distributions. This perturbation takes into account that the photos in the test set belong to the angles in between those of the training set. In particular, to sample an observation from class c and angle $d°$, we 'morph' the image with either the image of angle $d + 10°$ or the image of angle $d - 10°$. The morphed image vector v_{sample} is determined on the basis of a random number $m \in [-0.5, 0.5]$ as follows.

$$v_{\text{sample}} = \begin{cases} (1-|m|)v_d + |m|v_{d+10} & \text{, if } m > 0 \\ (1-|m|)v_d + |m|v_{d-10} & \text{, if } m \leq 0 \end{cases} \tag{16}$$

Figure 3 shows three real images and two morphed images of a sneaker viewed from angles $d = 20$, 30, and 40. The images indicated with $m = -0.5$ and $m = 0.5$ are morphed images of the image with $d = 30$.

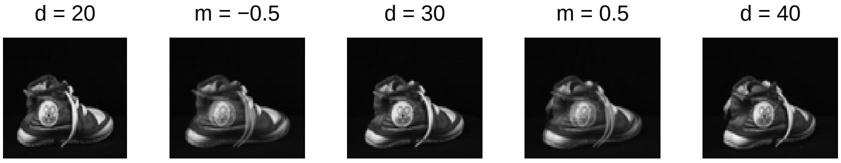

Figure 3: The left image is taken from angle $d = 20$, the middle image from $d = 30$, and the right image from $d = 40$. The images indicated with $m = -0.5$ and $m = 0.5$ are morphed images.

5.3 Training the Active Vision Models

In this subsection, we discuss the training of the active vision models EL and BH. As mentioned before, MI does not have a training procedure. We have fully described the training procedure of MB in Section 4.3.

We use an evolutionary algorithm to train the active vision models EL and BH. We start evolution by randomly initialising N different policies, $N = 20$. Each policy $\Pi : p(C \mid \mathbf{o}_i, \mathbf{a}_i) \rightarrow A$ is implemented by a fully connected multilayer feedforward network, with weights in the range [-1, 1]. The number of hidden neurons is half the number of inputs (classes), i.e., $\lfloor \frac{n}{2} \rfloor$. The genome of an individual policy is a vector of double values in the aforementioned range. The neural network transforms the current belief state into outputs that represent the possible actions. Therefore, it has n input neurons, where n is the number of objects under consideration, and it has 36 output neurons, i.e., the number of training or test angles (possible actions). The model executes the action corresponding to the output neuron with the highest activation value.

Training observations are generated according to the learned Gaussian distributions that represent $p(O|a,c)$, but we multiply the standard deviations of these distributions by a constant factor β in order to obtain interesting training cases that require actions. Figure 4 illustrates why we employ this factor β. The solid and dotted line illustrate the observation probability distributions of two different objects, projected on the first principal component. Clearly, if we sample according to these probability distributions, each observation can be classified correctly without the need for further actions. However, if we employ the factor β, we obtain observations that are more interesting. For example, the model can encounter observations that fall in between the probability distributions of the two different objects. The dashed line in the figure illustrates the way in which we sample observations for training. We set β to 4, on the basis of preliminary experiments on the tuning set of objects.

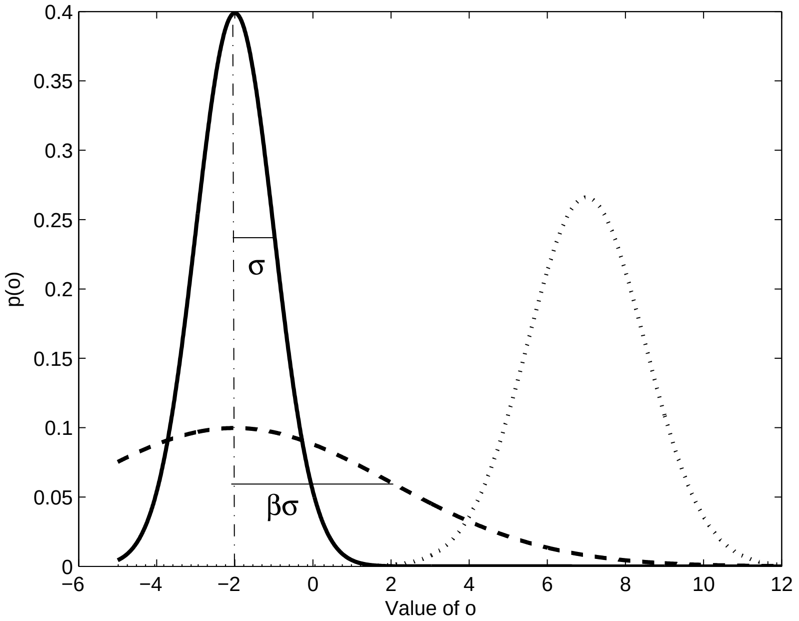

Figure 4: Illustration of the use of β. The solid and dotted line illustrate the observation probability distributions for the same action of two different objects, whose images are projected on the first principal component. The dashed line illustrates the way in which we sample observations for training if the model has to classify the object corresponding to the solid line, with $\beta = 4$.

To evaluate a policy, we apply it n times to the training set, so that it is evaluated once for each object. For BH we define the fitness of a policy as the proportion of correct classifications. For EL we define the fitness as the mean entropy-loss in the belief state. After all policies have been assigned a fitness, the best 5 policies are selected to form the next generation of policies. Per selected policy, we create 4 offspring by copying its genome 4 times. Then we apply mutation to the genomes of the offspring, where every gene (weight of the neural network) has a probability of p_m of being mutated. In our experiments, we set p_m to $\frac{1}{25}$. We mutate a gene by assigning it a random number from the interval $[-1, 1]$. This process of fitness evaluation, selection, and procreation, continues for $2n$ generations. Preliminary experiments on the tuning set have shown that this number of generations allows convergence of the evolutionary algorithm. The best policy of the last generation is returned as the trained policy. Since the outcome of training depends on the initial population, we perform three evolutions and select the policy that obtained the highest fitness. We evaluate this selected policy on the test set.

5.4 Testing the Active Vision Models

In total we perform fourteen experiments to compare the active vision models. These experiments are different in the number of objects that have to be classified, in the number of principal components, and in whether the model can determine the first viewing angle or not. Table 1 shows what experiments we perform. PC stands for principal component(s), n is the object subset size. In the experiments with a random first view angle, we employ two strategies to cope with the first observation: a conservative and a confident strategy. These strategies are further explained in Subsection 6.2.

	Model Determines First View Angle
1 PC	$n = 25, 50, 75, 100$
2 PC	$n = 25, 50, 75, 100$
	Random First View Angle
1 PC	$n = 25, 50, 75$ - conservative strategy
1 PC	$n = 25, 50, 75$ - confident strategy

Table 1: The fourteen experiments that we perform to compare the four active vision models. PC stands for principal component(s), n for the object subset size. For the experiments with a random first view angle, we employ a conservative or a confident strategy to cope with the first observation.

The results of each experiment are obtained by performing multiple experimental runs. These experimental runs start by randomly selecting an object subset from the ALOI-database and by learning the observation probability distributions. Then, each active vision model is trained on the training angles (if necessary) and tested on the test angles. Every model encounters the same test objects and, in the case that the model does not determine the first viewing angle, the same first test viewing angles. Since we regarded the experiments with one principal component in which the models determine the first viewing angle as the most standard experiments, we performed 50 different experimental runs to obtain the results for these experiments. For the other experiments we relied on 30 different experimental runs.

6 Results

In this section, we first show the results on the task where the first view angle is determined by the models. Then we proceed with the results of the experiments in which the first view angle is picked at random and is unknown to the models.

6.1 Model Determines First View Angle

Table 2 shows the mean error and standard error of the mean of all active vision models for the experiments with one principal component. The results of each active vision model are shown in a column, while rows represent the experiments for a specific number of objects.

	MI	EL	MB	BH	RA	PA
25	84.6 (1.3)	86.1 (1.1)	81.8 (1.3)	**87.0 (1.0)**	74.4 (0.1)	68.6 (0.2)
50	**73.6 (1.2)**	71.0 (1.0)	71.1 (1.0)	72.1 (1.0)	64.0 (1.0)	47.9 (1.1)
75	63.2 (0.9)	61.4 (0.9)	62.2 (0.9)	**64.7 (1.0)**	58.4 (0.9)	34.2 (1.0)
100	56.8 (0.9)	54.3 (0.9)	55.6 (0.9)	**62.2 (1.3)**	58.5 (0.6)	27.9 (0.9)

Table 2: One principal component: results of experiments on the test set. Average performance and standard error of the mean per active vision model, based on 50 experimental runs.

The table shows that for 25, 50, and 75 objects, all methods have a higher mean performance than the random active vision strategy, confirming the results from the literature [3, 10, 16]. However, it also shows that the random strategy deteriorates less with a growing object subset. For 100 objects, the random strategy performs better than all models, except BH. Figure 5 shows the effects of enlarging the object subset on the performances of the active vision models. Figure 6 to 9 more clearly bring out the differences between the active vision models.

Figure 6 to 9 show that model MB performs worst on the smallest subset, but performs better than model EL for the bigger subsets with an increasing difference. The active vision model BH performs better than model EL on all subset sizes, with an increasing difference. It attains the highest performance on three of the four subsets, and is the only active vision model that outperforms the random active vision strategy for 100 objects. A last observation is that MI and BH always perform better than MB. Table 3 shows what performance differences are statistically significant. Statistical significance was determined with a randomisation test [5], with $p < 0.05$. In the randomisation test, we determined the probability with which the performance difference between two

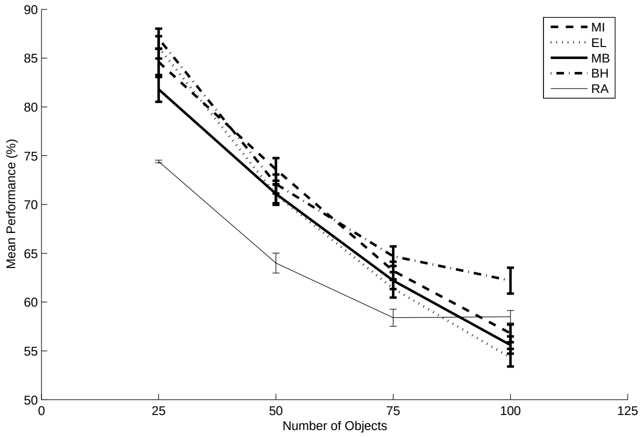

Figure 5: One principal component: mean performance and standard error of the mean of all active vision models except PA for subset sizes 25, 50, 75, and 100.

models is equal to, or bigger than, the experimental difference, given that the two models perform equally well. In order to determine this probability, we randomly divided the experimental results in two sets for a thousand times, and counted the number of times for which the difference in average performance of the two sets was equal to or bigger than the original difference. If this number of times was smaller than 50, the assumption that the models perform equally well is discarded: the results are then regarded to be statistically significant. In each cell of the table we include the numbers of the object subset sizes for which the row's model is significantly better than the column's model. For example, EL significantly outperforms MB for a subset size of 25. Therefore, we included 25 in the cell with row EL and column MB.

	MI	EL	MB	BH	RA	PA
MI					25, 50, 75	25, 50, 75, 100
EL			25		25, 50, 75	25, 50, 75, 100
MB					25, 50, 75	25, 50, 75, 100
BH	100	75, 100	25, 100		25, 50, 75, 100	25, 50, 75, 100
RA		100	100			25, 50, 75, 100
PA						

Table 3: One principal component: statistically significant results for the experiments with $n = 25, 50, 75$, and 100 ($p < 0.05$). If a subset size is mentioned in a cell, then the active vision model of the cell's row significantly outperforms the model of the cell's column.

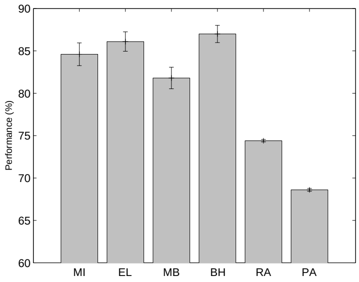

Figure 6: 25 objects: mean performances and standard errors of the mean.

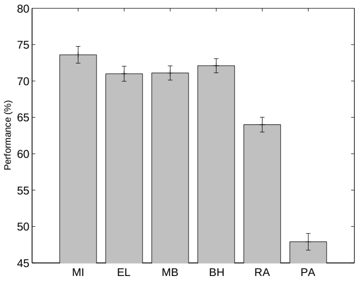

Figure 7: 50 objects: mean performances and standard errors of the mean.

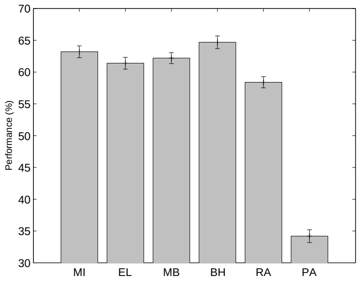

Figure 8: 75 objects: mean performances and standard errors of the mean.

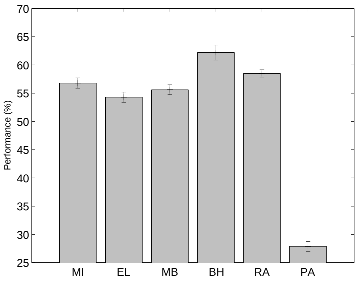

Figure 9: 100 objects: mean performances and standard errors of the mean.

Finally, we show how many actions the different active vision models take to classify an object. Table 4 shows per subset size and per active vision model, how many actions it takes on average to classify an object (first number in each cell). We also recorded for each experimental run the maximum number of actions taken by each active vision model. The table shows what the maximum number of actions is that a model takes, averaged over all experimental runs (second number in each cell). We observe from the table that on average RA performs the most actions, with as only exception the subset size of 75 objects. Between the other active vision models, there do not seem to be big differences in the number of actions that they take. The average number of actions for all models and all subsets is close to 1, indicating that the first action often already leads to the classification of the object under evaluation.

Table 5 shows the outcome of the experiments with two principal components. It mainly shows that the objects in the ALOI-database are rather dissimilar, since the addition of one extra principal component causes a ceiling

	MI	EL	MB	BH	RA	PA
25	1.06 / 1.82	1.04 / 1.68	1.04 / 1.64	1.05 / 1.74	1.19 / 2.26	1 / 1
50	1.09 / 2.00	1.08 / 2.08	1.08 / 2.08	1.08 / 2.00	1.30 / 2.67	1 / 1
75	1.12 / 2.14	1.11 / 2.22	1.17 / 2.44	1.18 / 2.38	1.15 / 3.06	1 / 1
100	1.15 / 2.10	1.16 / 2.40	1.22 / 2.74	1.35 / 2.54	1.52 / 3.20	1 / 1

Table 4: One principal component: number of actions performed per active vision model, based on 50 experimental runs. The first number in each cell is the average number of actions performed per run. The second number is the average of the maximal number of actions performed for one object in an object subset.

effect for the active vision model performances. The performance of the passive vision model PA is rather good. As the subset size increases, the active vision models outperform more and more RA and PA. Table 6 shows the statistical significance of the performance differences. Table 7 shows the average number of actions and the average maximum number of actions of all active vision models for the task with two principal components.

	MI	EL	MB	BH	RA	PA
25	99.6 (0.3)	**99.7 (0.3)**	99.3 (0.3)	**99.7 (0.3)**	95.1 (0.9)	97.9 (0.5)
50	**98.5 (0.4)**	98.3 (0.4)	97.7 (0.4)	**98.5 (0.3)**	92.2 (0.6)	92.2 (0.8)
75	97.3 (0.4)	**97.7 (0.3)**	97.3 (0.4)	97.6 (0.3)	91.2 (0.6)	88.5 (0.7)
100	95.8 (0.6)	**96.2 (0.5)**	95.7 (0.6)	96.1 (0.5)	86.0 (0.5)	86.0 (0.7)

Table 5: Two principal components: results of experiments on the test set. Average performance and standard error of the mean per active vision model, based on 30 experimental runs.

	MI	EL	MB	BH	RA	PA
MI					25, 50, 75, 100	25, 50, 75, 100
EL					25, 50, 75, 100	25, 50, 75, 100
MB					25, 50, 75, 100	25, 50, 75, 100
BH					25, 50, 75, 100	25, 50, 75, 100
RA						75
PA					25	

Table 6: Two principal components: statistically significant results for the experiments with $n = 25, 50, 75$, and 100 ($p < 0.05$). If a subset size is mentioned in a cell, then the active vision model of the cell's row significantly outperforms the model of the cell's column.

	MI	EL	MB	BH	RA	PA
25	1.00 / 1.03	1.00 / 1.00	1.00 / 1.00	1.00 / 1.00	1.01 / 1.13	1 / 1
50	1.00 / 1.20	1.00 / 1.10	1.00 / 1.13	1.00 / 1.10	1.02 / 1.60	1 / 1
75	1.01 / 1.47	1.01 / 1.30	1.01 / 1.33	1.00 / 1.23	1.04 / 1.93	1 / 1
100	1.01 / 1.50	1.00 / 1.37	1.00 / 1.40	1.00 / 1.37	1.04 / 2.03	1 / 1

Table 7: Two principal components: number of actions performed per active vision model, based on 30 experimental runs. The first number in each cell is the average number of actions performed per run. The second number is the average of the maximal number of actions performed for one object in an object subset.

6.2 Random First View Angle

In this subsection, we show the results for the experiments in which the first view angle is picked at random. Since this angle is not known to the models, they have to follow a different strategy for updating the belief state for the first observation, than for subsequent observations. In our experiments, we tried out two different strategies: a conservative and a confident strategy. Both strategies take the observation o and determine for each action-class pair what the probability is that it generated the observation. The conservative strategy then excludes the classes that have zero probability of generating the observation, assigning equal probabilities to the remaining classes. The confident strategy sums the probabilities over all classes per action. Then it assumes that the action taken is the one with the highest sum of probabilities and updates the belief state with the help of o and this most probable action.

Because the experiments with an unknown first action took more time, we only performed experiments for $n = 25, 50$, and 75. Table 8 shows the results for the task in which the first view is picked at random, where the agent employs the conservative strategy to handle the first observation. We see that the performance of PA is much lower than for the experiments shown in Table 2. The reason for this is that it is much harder to immediately classify an object on the basis of one observation, if the angle is unknown. Many of the other performances in the table are higher than those in Table 2. The reason for this becomes evident, when investigating the number of actions executed by the active vision models, shown in Table 10. The models take on average roughly one action more when they do not determine the first angle. This one action is the random, unknown angle. Apparently, the first observation leads to the exclusion of some objects, but does not often lead to an immediate classification. With one observation more, it is not surprising that the performances are higher than for the experiment in Section 6.1. The only active vision model that does not perform better, is BH. Concerning the comparison of the active vision models, the main observation is that MI performs better than the other models on all three subset sizes. Table 9 shows the statistical significance of the performance differences between the various active vision models.

Table 11 shows the results for the confident strategy. Clearly, the performances obtained are lower than for the conservative strategy. Comparing Table 13 with Table 10 reveals that, on average, the active vision models take less actions for the confident strategy than for the conservative strategy. Assuming that the random action corresponds to the most probable action results in quicker, but faultier classifications. Table 11 shows that BH and MI perform best for this strategy. Table 12 shows the performance differences that were statistically significant.

	MI	EL	MB	BH	RA	PA
25	**89.3 (1.1)**	87.0 (1.2)	80.2 (1.5)	87.0 (1.2)	76.0 (1.0)	21.8 (1.4)
50	**75.0 (1.2)**	73.5 (1.3)	72.5 (1.4)	72.0 (1.2)	65.0 (0.7)	12.1 (1.3)
75	**66.9 (1.1)**	63.0 (1.3)	66.2 (0.9)	63.8 (1.2)	60.3 (0.5)	9.3 (0.9)

Table 8: Conservative strategy: results of experiments with one principal component on the test set where the first view is selected at random. Average performance and standard error of the mean per active vision model, based on 30 experimental runs.

	MI	EL	MB	BH	RA	PA
MI		75	25		25, 50, 75	25, 50, 75
EL			25		25, 50, 75	25, 50, 75
MB		75			25, 50, 75	25, 50, 75
BH			25		25, 50, 75	25, 50, 75
RA						25, 50, 75
PA						

Table 9: Conservative strategy: statistically significant results for the experiments with one principal component and random first angle for $n = 25, 50, 75$ ($p < 0.05$). If a subset size is mentioned in a cell, then the active vision model of the cell's row significantly outperforms the model of the cell's column.

	MI	EL	MB	BH	RA	PA
25	2.04 / 3.03	2.04 / 2.73	2.09 / 2.90	2.04 / 2.60	2.16 / 3.87	1 / 1
50	2.08 / 3.10	2.07 / 3.10	2.16 / 3.30	2.08 / 3.17	2.29 / 4.37	1 / 1
75	2.13 / 3.27	2.12 / 3.20	2.27 / 3.57	2.13 / 3.07	2.41 / 4.83	1 / 1

Table 10: Conservative strategy: number of actions performed per active vision model, based on 30 experimental runs. The first number in each cell is the average number of actions performed per run. The second number is the average of the maximal number of actions performed for one object in an object subset.

	MI	EL	MB	BH	RA	PA
25	49.9 (1.2)	43.1 (1.3)	47.8 (1.1)	**50.2 (1.4)**	39.9 (0.8)	23.6 (1.8)
50	**51.3 (0.8)**	46.3 (1.2)	47.3 (0.9)	51.2 (0.9)	37.0 (0.5)	13.0 (1.2)
75	47.6 (0.8)	46.4 (1.0)	41.9 (0.7)	**48.1 (0.8)**	32.3 (0.6)	7.2 (0.7)

Table 11: Confident strategy: results of experiments with one principal component on the test set where the first view is selected at random. Average performance and standard error of the mean per active vision model, based on 30 experimental runs.

	MI	EL	MB	BH	RA	PA
MI		25, 50	50, 75		25, 50, 75	25, 50, 75
EL			75		25, 50, 75	25, 50, 75
MB		25			25, 50, 75	25, 50, 75
BH		25, 50	50, 75		25, 50, 75	25,50, 75
RA						25, 50, 75
PA						

Table 12: Confident strategy: statistically significant results for the experiments with one principal component and random first angle for $n = 25, 50, 75$ ($p < 0.05$). If a subset size is mentioned in a cell, then the active vision model of the cell's row significantly outperforms the model of the cell's column.

	MI	EL	MB	BH	RA	PA
25	1.57 / 3.03	1.57 / 3.27	1.57 / 3.17	1.59 / 3.23	1.57 / 3.23	1 / 1
50	1.86 / 3.23	1.88 / 3.67	1.86 / 3.30	1.89 / 3.60	1.87 / 3.73	1 / 1
75	1.95 / 3.27	1.97 / 3.62	1.96 / 3.62	1.97 / 3.96	1.97 / 4.00	1 / 1

Table 13: Confident strategy: number of actions performed per active vision model, based on 30 experimental runs. The first number in each cell is the average number of actions performed per run. The second number is the average of the maximal number of actions performed for one object in an object subset.

7 Analysis

In this section, we perform an analysis of the most important experimental results. Table 14 is a summary of the results of all fourteen experiments concerning the performance differences between the models (Section 6). The first number shows for each model in how many experiments it had a higher mean performance than each other model. The second number shows how many times this difference was statistically significant. The table shows that MI generally performs better than both EL and MB. We perform a small theoretical analysis that provides theoretical support for this result in subsection 7.1. In addition, the table shows that BH also generally performs better than EL and MB. We discuss why BH outperforms EL in subsection 7.2. Then, we investigate remaining comparisons between the models in subsection 7.3. Finally, in subsection 7.4, we provide an explanation for the surprising result that RA outperformed other active vision models in the experiment with one principal component, where the first view is determined by the model.

	MI	EL	MB	BH	RA	PA
MI		10 / 3	13 / 3	5 / 0	13 / 13	14 / 14
EL	4 / 0		8 / 3	3 / 0	13 / 13	14 / 14
MB	0 / 0	6 / 2		2 / 0	13 / 13	14 / 14
BH	8 / 1	9 / 4	12 / 5		14 / 14	14 / 14
RA	1 / 0	1 / 1	1 / 1	0 / 0		11 / 11
PA	0 / 0	0 / 0	0 / 0	0 / 0	1 / 1	

Table 14: Summary of performance differences of all experiments. The first number indicates the number of times the model of the row had a higher mean performance than the model of the column. The second number indicates how many times this higher performance was statistically significant in itself.

7.1 Comparison of MI with EL and MB

The active vision model MI selects an action a^*, according to[1]:

$$a^* = \text{argmax}_{a_i} E[H(C \mid \mathbf{o}_{i-1}, \mathbf{a}_{i-1}) - H(C \mid \mathbf{o}_i, \mathbf{a}_i)] \tag{17}$$

MI calculates the expected entropy loss in the belief state (i.e., mutual information), given the belief state and the probability distribution of observations O for an action a. MI maximises:

$$E[H(C \mid \mathbf{o}_{i-1}, \mathbf{a}_{i-1}) - H(C \mid \mathbf{o}_{i-1}, \mathbf{a}_{i-1}, a, o)] = \tag{18}$$

[1] We formulate the action selection of MI in a slightly different manner than in Section 4 in order to better compare it with EL's action selection. Note however, that Equation 19 is equal to Equation 10.

$$\sum_{c \in C} \sum_{o \in O} p(c \mid \mathbf{o}_{i-1}, \mathbf{a}_{i-1}, a) p(o \mid c, a) \log\left(\frac{p(o \mid c, a)}{\sum_{c \in C} p(c \mid \mathbf{o}_{i-1}, \mathbf{a}_{i-1}) p(o \mid c, a)}\right) \quad (19)$$

On the contrary, EL selects actions $\mathbf{a}_t^*$, according to:

$$\mathbf{a}_t^* = \mathrm{argmax}_{\mathbf{a}_t} E[H(C) - H(C \mid \mathbf{o}_t, \mathbf{a}_t)] \quad (20)$$

Where t is the time step at which a class is assigned to the object under consideration. Comparing Equation 20 with Equation 17 shows that EL maximises the entropy reduction over multiple time steps, while MI maximises the entropy reduction over one time step. As a consequence, an advantage of EL is that it can perform non-greedy action selection while MI always performs greedy action selection. However, in the current object classification task, this advantage does not seem to be of high importance. The main reason for this is that the active vision models can change the view angle to any other angle at any time step. Therefore, the active vision models do not need to employ a sequence of non-greedy actions to arrive at a discriminative view angle. In addition, the experimental results show that classification can often already be performed with one or two observations. This does not leave much room for employing non-greedy action selection.

The reason that MI outperforms EL in many experiments seems to be the manner of action selection. Where MI calculates the expected entropy reduction at every time step, EL learns an action mapping on the basis of experience. During training it performs an action a, which leads to a specific observation o, and to an entropy loss in the belief state:

$$H(C \mid \mathbf{o}_{i-1}, \mathbf{a}_{i-1}) - H(C \mid \mathbf{o}_{i-1}, \mathbf{a}_{i-1}, a, o) = \quad (21)$$

$$\sum_{c \in C} p(c \mid \mathbf{o}_{i-1}, \mathbf{a}_{i-1}, a) \log\left(\frac{p(o \mid c, a)}{\sum_{c \in C} p(c \mid \mathbf{o}_{i-1}, \mathbf{a}_{i-1}) p(o \mid c, a)}\right) \quad (22)$$

The difference between Equation 22 and Equation 19 is that MI sums over $p(O)$ to calculate the expected entropy loss. On the contrary, the action mapping of EL averages over the observations received during training. For $\beta = 1$, EL receives observations according to the distribution $p(O \mid c, a)$ during training. With an unlimited amount of training time EL should arrive at the same actions as MI, when the same belief state is given. However, in our implementation, EL has a limited amount of training time. In addition, the neural network that implements Π generalises over the input space $p(C \mid \mathbf{o}_{i-1}, \mathbf{a}_{i-1})$. Therefore, similar belief states are mapped to similar actions, even if using $E[H(C \mid \mathbf{o}_{i-1}, \mathbf{a}_{i-1}) - H(C \mid \mathbf{o}_{i-1}, \mathbf{a}_{i-1}, a, o)]$ suggested a different action. We believe that this is the reason that MI outperforms EL on most experimental settings. One could argue that the factor $\beta = 4$ is the reason for EL's worse performance, since it prevents EL from learning the real observation probability distribution. However, experiments on the tuning set showed that setting β to 4 improved performance.

MB uses Equation 12 and 14 to estimate $E[H(C)-H(C|a)|k]$. This estimate makes use of a very coarse generalisation over the input space by only considering the mode of the belief state. The amount of resulting input states is equal to the number of classes, $|C|$. In addition, this estimate does not take into account past observations and actions. The rough generalisation over the belief state and the neglect of past information has as a result that MI outperforms MB.

7.2 Comparison of BH and EL

As mentioned in Subsection 4.4, the only difference between BH and EL is the fitness function. EL is trained on the basis of the entropy reduction in the belief state, while BH is trained on the basis of classification performance. In other words, EL's action policy is optimised so that the model is very sure about its classification, while BH's action policy is optimised so that the model's classification is correct.

In a practical setting, belief state updates on the basis of the estimated probability distributions might not always be correct. Therefore, some belief states might be misleading, resulting in suboptimal actions and possibly wrong classifications. None of the probabilistic active vision models has a means for recovering from such a belief state. BH *can* recover from such a misleading belief state, e.g., by selecting an action that does not lead to the maximal entropy reduction but that will lead to correct classification. We think that for this reason BH outperforms EL on most experimental settings.

7.3 Other Comparisons

We did not yet compare BH with MI and MB. Theoretically, BH has the advantage over the other two models that it can recover from misleading belief states. However, MI has the advantage over BH that it does not generalise over the belief state with a neural network. Therefore, it is difficult to determine theoretically whether BH should outperform MI or vice versa. The experimental results seem to indicate that this depends on the type of problem. BH outperforms MI when the model can determine the first action, but MI outperforms BH when the first angle is randomly chosen and unknown to the model. The experimental results are clearer when it concerns the comparison between BH and MB. In almost all experiments, BH outperforms MB.

7.4 Subset Size and Random Active Vision Strategy

In Section 6, we noted that the performance difference between the active vision models and the random active vision strategy decreases, as the number of objects increases. For 100 objects, the random strategy even outperforms most of the active vision models. There seem to be two reasons for this.

The first reason for RA's good performance seems to lie in the number of actions that it takes. As stated in Section 6.1, the random active vision model takes more actions on average than the other models. The second reason for RA's good performance might be that, depending on the object subset, it becomes easier to select discriminative views. When the number of objects increases, the classification task becomes more difficult. Namely, there are more different subgroups of objects that can appear similar from certain angles while they appear dissimilar from others. However, this also implies that there are more angles that can disambiguate different subgroups of objects and that it becomes easier to find such an angle.

8 Discussion and Conclusions

In this report, we have described four active vision models in a common framework and we compared them to each other empirically. In this section, we first indicate the limitations of our empirical comparison of the active vision models. Then we discuss the insights obtained on the difference between the probabilistic and the behavioural approach to active vision. Subsequently, we discuss the findings for the differences within the probabilistic approach. Finally, we discuss what the experiments reveal on the relation between the usefulness of active vision, the object subset size, and the richness of input features.

One limitation of our comparison of the different active vision models is that we compared the active vision models on a view-based three-dimensional object classification task only. This type of task, although widely used in the literature (e.g., [10, 16]), has specific properties that may not be present in many real-world problems. For example, in our task, actions represent absolute angles. As a result, any angle view can be accessed at any time step. In real-world problems, this might not be the case: an action might represent a shift in the angle and larger shifts may require multiple actions. If this were the case, then it would be wiser to employ another belief state update (e.g., a Naive Bayesian one as in [16]) and methods that are able to find non-greedy action selection strategies would perhaps be at an advantage (e.g., EL and BH). In addition, in real-world problems there might be many more possible actions than just changing the angle from which the model views an object. Additional actions, and a continuous instead of discrete action space might also change the problem significantly, and thus influence the performance differences between the various active vision models.

Even though our comparison has its limits, the experimental results give some insight into the differences between behavioural and probabilistic active vision models. For example, one could expect a behavioural model to perform worse on our object classification task, since the action selection strategy is not well formalised in terms of reducing entropy. However, the results show that BH is not at a disadvantage. The model BH has a high performance in most experiments compared to the other active vision models, especially in the experiments where the model can select the first viewing angle. Importantly, it outperforms

EL in most experiments, while in general not performing more actions. The lack of a well formalised action selection strategy does not necessarily result in worse performances. However, one can argue that the model BH is not yet a very 'behavioural' model. Specifically, it incorporates an explicit belief state, which is not typical for the behavioural approach to active vision. Therefore, future comparisons should incorporate active vision models without belief states as well.

The experiments also provide insight into the differences within the probabilistic approach to active vision. Our results and analysis indicate that in most cases MI outperforms both EL and MB, while it does not on average take more actions than those models. Of course, performance is not the only factor of importance in choosing between these three models. In Section 4 we also discussed differences in training and execution time. For example, MI has no training time besides the estimation of the observation distributions, and has a long execution time. On the contrary, EL has a long training time, but a very short execution time. MB has both a short training and execution time and does not seem to perform much worse than EL.

The experimental results reveal a relation between the usefulness of active vision, the number of objects, and the richness of input features. The experiments confirm two general expectations on this relation, but also indicate two challenges for the field of active vision. For example, as expected, performance on the classification task improves if input features (in our case principal components) are added. However, for the object subset sizes that we studied, this improvement was much larger than expected. In particular, the addition of the second principal component improved the performance on the classification task more than the addition of sensible action selection. The mean performance of the passive vision model PA in Table 5 is higher than the mean performances of the active vision models in Table 2. This result introduces a doubt regarding the usefulness of active vision in the light of a trade-off with better or more input features. Of course, there can be cases in which a passive vision strategy cannot solve a classification problem. For example, in [10], 3-D puppets are used that are only different when seen from one particular viewing angle. However, it is to be seen whether many real-world problems are of this type. Therefore, it seems to be a challenge for the field of active vision to find action selection strategies that also improve the performance when visual information has already been optimised for the task. Another general expectation that the results confirmed is that performance should go down, as the object subset grows. Surprisingly however, under some conditions the performance difference between RA and other active vision models becomes smaller with increasing object subset size. In the experiment where the model can select its first view angle and with a subset size of 100 objects, RA even outperforms all other active vision models, except for BH. This result suggests that the challenge in applying active vision models to large object sets does not only reside in dealing with computation time (increasingly long training or execution times), but also with sensible action selection.

References

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

[2] Thomas Bäck. *Evolutionary Algorithms in Theory and Practice.* Oxford University Press, New York, Oxford, 1996.

[3] H. Borotschnig, L. Paletta, M. Prantl, and A. Pintz. Appearance-based active object recognition. *Image and Vision Computing*, 18:715–727, 2000.

[4] H. Borotschnig, L. Paletta, M. Prantl, and A. Pinz. A comparison of probabilistic, possibilistic and evidence theoretic fusion schemes for active object recognition. *Computing*, 62:293–319, 1999.

[5] P. Cohen. *Empirical Methods for Artificial Intelligence.* MIT Press, Cambridge, Massachusetts, 1995.

[6] G. de Croon, E.O. Postma, and H.J. van den Herik. A situated model for sensory-motor coordination in gaze control. *Pattern Recognition Letters*, 27:1181–1190, 2006.

[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*, 13(3):227–242, 2005.

[8] F. Deinzer, J. Denzler, and H. Niemann. Viewpoint selection - planning optimal sequences of views for object recognition. In N. Petkov and M.A. Westenberg, editors, *Computer Analysis of Images and Patterns, 10th International Conference, CAIP 2003, Groningen, The Netherlands*, pages 65–73. Springer, 2003.

[9] J. Denzler and C.M. Brown. Optimal selection of camera parameters for state estimation of static systems: an information theoretic approach. *Technical Report, Computer Science Department, University of Rochester*, 2000.

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

[11] B. Deutsch, M. Zobel, J. Denzler, and H. Niemann. Multi-step entropy based sensors control for visual object tracking. In *Pattern Recgonition, 26^{th} DAGM Symposium*, pages 359–366, Berlin, 2004. Springer-Verlag.

[12] D. Floreano, T. Kato, D. Marocco, and E. Sauser. Coevolution of active vision and feature selection. *Biological Cybernetics*, 90:3:218–228, 2004.

[13] J.M. Geusebroek, G.J. Burghouts, and A.W.M. Smeulders. The Amsterdam library of object images. *International Journal of Computer Vision*, 61:103–112, 2005.

[14] B. J. A. Kröse and R. Bunschoten. Probabilistic localization by appearance models and active vision. In *IEEE Int. Conf. on Robotics and Automation*, pages 2255–2260, 1999.

[15] H. Murase and S.K. Nayar. Visual learning and recognition of 3-D objects from appearance. *International Journal of Computer Vision*, 14:5–24, 1995.

[16] L. Paletta, M. Prantl, and A. Pinz. Reinforcement learning for autonomous three-dimensional object recognition. In *Proc. 6th Symposium on Intelligent Robotics Systems*, Edinburgh, UK, 1998.

[17] M. Partridge and R.A. Calvo. Fast dimensionality reduction and simple PCA. *Intelligent Data Analysis*, 2:203–214, 1998.

[18] B. Schiele and J.L. Crowley. Transinformation of object recognition and its application to viewpoint planning. *Robotics and Autonomous Systems*, 21:95–106, 1997.

[19] B. Schiele and J.L. Crowley. Transinformation for active object recognition. In *Proceedings of the Sixth International Conference on Computer Vision (ICCV)*, pages 249–255, Washington DC, USA, 1998. IEEE Computer Society.

[20] R.S. Sutton and A.G. Barto. *Reinforcement Learning: An Introduction.* MIT Press, Cambridge, MA, A Bradford Book, 1998.