You Only Look as Much as You Have To

Using the Free Energy Principle for Active Vision

Toon Van de Maele[1], Tim Verbelen[1], Ozan Çatal[1], Cedric De Boom[1], and Bart Dhoedt[1]

IDLab, Department of Information Technology
Ghent University - imec
Ghent, Belgium
toon.vandemaele@ugent.be

Abstract. Active vision considers the problem of choosing the optimal next viewpoint from which an autonomous agent can observe its environment. In this paper, we propose to use the active inference paradigm as a natural solution to this problem, and evaluate this on a realistic scenario with a robot manipulator. We tackle this problem using a generative model that was learned unsupervised purely from pixel-based observations. We show that our agent exhibits information-seeking behavior, choosing viewpoints of regions it has not yet observed. We also show that goal-seeking behavior emerges when the agent has to reach a target goal, and it does so more efficiently than a systematic grid search.

Keywords: Active Vision, Active Inference, Deep Generative Modelling, Robotic planning

1 Introduction

Active vision considers an observer that can act by controlling the geometric properties of the sensor in order to improve the quality of the perceptual results [1]. This problem becomes apparent when considering occlusions, a limited field of view or a limited resolution of the used sensor [2]. In many cases, selecting the next viewpoint should be done as efficiently as possible due to limited resources for processing the new observations and the time it takes to reach the new observation pose. This problem is traditionally solved with frontier-based methods [15] in which the environment is represented as an occupancy grid. These approaches rely on evaluating engineered utility functions that estimate the amount of new information provided for all potential viewpoints [15, 8]. Usually this utility function represents the amount of unobserved voxels that a given viewpoint will uncover. Instead of using hand-crafted heuristics, this function can also be learned from data [8, 9]. A different approach is to predict the optimal viewpoint with respect to reducing uncertainty and ambiguity directly from a reconstructed volumetric grid [3, 10]. A different bio-inspired method for active vision is proposed by Rasouli et. al. [13] in which the action is driven by a

visual attention mechanism in conjunction with a non-myopic decision-making algorithm that takes previous observations at different locations in account.

Friston et. al. [11, 7] cast the active vision problem as a low dimensional, discrete state-space Markov decision process (MDP) that can be solved using the active inference framework. In this paradigm, agents act in order to minimize their surprise, i.e. their free energy. In this paper, instead of using an explicit 3D representation, or a simple MDP formulation of the environment, we learn a generative model and latent state distribution purely from observations. Previous work also used deep learning techniques to learn the generative model in order to engage in active inference [16], while other work has created an end-to-end active inference pipeline using pixel-based observations [14]. Similar to Friston et al. [6, 11, 7], we then use the expected free energy to drive action selection. Similar to the work of Nair, Pong et. al [12] where the imagined latent state is used to compute the reward value for optimizing reinforcement learning tasks and the work of Finn and Levine [5], where a predictive model is used that estimates the pixel observations for different control policies, we employ the imagined observations form the generative model to compute the expected free energy. We evaluate our method on a grasping task with a robotic manipulator with an in-hand camera. In this task, we want the robot to get to the target object as fast as possible. For this reason we consider the case of best viewpoint selection. We show how active inference yields information-seeking behavior, and how the robot is able to reach goals faster than random or systematic grid search.

2 Active Inference

Active inference posits that all living organisms minimize free energy (FE) [6]. The variational free energy is given by:

$$\begin{aligned} F &= \mathbb{E}_Q[\log Q(\tilde{\mathbf{s}}) - \log P(\tilde{\mathbf{o}}, \tilde{\mathbf{s}}, \pi)] \\ &= \mathrm{D}_{KL}[Q(\tilde{\mathbf{s}})||P(\tilde{\mathbf{s}}, \pi)] - \mathbb{E}_Q[\log P(\tilde{\mathbf{o}}|\tilde{\mathbf{s}}, \pi)], \end{aligned} \tag{1}$$

where $\tilde{\mathbf{o}}$ is a sequence of observations, $\tilde{\mathbf{s}}$ the sequence of corresponding model belief states, π the followed policy or sequence of actions taken, and $Q(\tilde{\mathbf{s}})$ the approximate posterior of the joint distribution $P(\tilde{\mathbf{o}}, \tilde{\mathbf{s}}, \pi)$. Crucially, in active inference, policies are selected that minimize the expected free energy $G(\pi, \tau)$ for future timesteps τ [6]:

$$G(\pi, \tau) \approx -\mathbb{E}_{Q(\mathbf{o}_\tau|\pi)}[\mathrm{D}_{KL}[Q(\mathbf{s}_\tau|\mathbf{o}_\tau, \pi)||Q(\mathbf{s}_\tau|\pi)]] - \mathbb{E}_{Q(\mathbf{o}_\tau|\pi)}[\log P(\mathbf{o}_\tau)]. \tag{2}$$

This can be viewed as a trade-off between an epistemic, uncertainty-reducing term and an instrumental, goal-seeking term. The epistemic term is the Kullback-Leibler divergence between the expected future belief over states when following policy π and observing $\mathbf{o}_\tau$ and the current belief. The goal-seeking term is the likelihood that the goal will be observed when following policy π.

3 Environment and approach

In this paper, we consider a simulated robot manipulator with an in-hand camera which can actively query observations from different viewpoints or poses by moving its gripper, as shown in Figure 1. The robotic agent acts in a static workspace, in which we randomly spawn a red, green and blue cube of fixed size. Each such configuration of random cube positions is dubbed a scene, and the goal of the robot is to find a cube of a particular color in the workspace. The agent initially has no knowledge about the object positions and has to infer this information from multiple observations at different poses. Example observations for different downward facing poses are given in Figure 2.

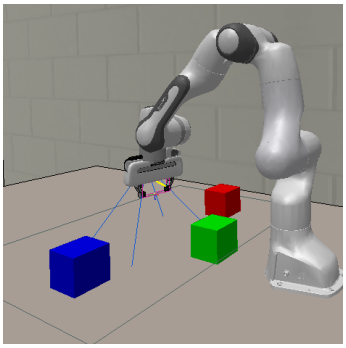


Fig. 1: Franka Emika Panda robot in the CoppeliaSim simulator in a random scene with three colored cubes.

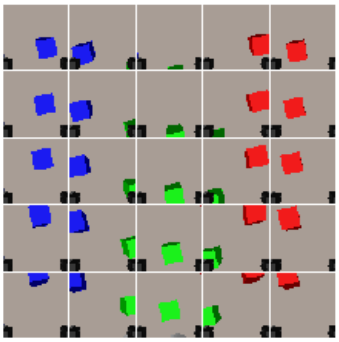


Fig. 2: Sampled observations on a grid of potential poses used for evaluating the expected free energy.

To engage in active inference, the agent needs to be equipped with a generative model. This generative model should be able to generate new observations given an action or in this particular case, the new robot pose. In contrast with [11, 7], we do not fix the generative model upfront, but learn it from data. We generate a dataset of 250 different scenes consisting of approximately 25 discrete time steps in which the robot observes the scene from a different viewpoint. Using this dataset we train two deep neural networks to approximate the likelihood distribution $P(\mathbf{o}_t|\mathbf{s}_t, \pi)$ and approximate posterior distribution $Q(\mathbf{s}_t|\mathbf{o}_t, \pi)$ as multivariate Gaussian distributions. In our notation, $\mathbf{o}_t$ and $\mathbf{s}_t$ respectively represent the observation and latent state at discrete timestep t. Both distributions are conditioned by the policy π, the action that the robot should take in order to acquire a new observation, or equivalently, the new observation viewpoint. The models are optimized by minimizing the free energy from Equation 1, with a zero-mean isotropic Gaussian prior $P(\mathbf{s}_t|\pi) = \mathcal{N}(0, 1)$. Hence the system is trained as an encoder-decoder to predict scene observations from unseen poses, given a number of observations from the same scene at different poses. This is similar to a Generative Query Network (GQN) [4]. For more details on the model architecture and training hyperparameters, we refer to Appendix A.

At inference time, the policy π, or equivalently the next observer pose, is selected by evaluating Equation (2) for a number of candidate policies and selecting the policy that evaluates to the lowest expected free energy. These candidate policies are selected by sampling a grid of poses over the workspace. The trained decoder extracts the imagined observation for each of the candidate policies and the state vector acquired through encoding the initial obsevations. The corresponding expected posterior distributions are computed by forwarding these imagined observations together with the initial observations through the encoder. For the goal-seeking term, we provide the robot with a preferred observation, i.e. the image of the colored cube to fetch, and we evaluate $\log P(\mathbf{o}_\tau)$. The epistemic term is evaluated by using the likelihood model to imagine what the robot would see from the candidate pose, and then calculating the KL divergence between the state distributions of the posterior model before and after "seeing" this imagined observation. The expectation terms are approximated by drawing a number of samples for each candidate pose.

4 Experiments

We evaluate our system in two scenarios. In the first scenario, only the epistemic term is taken into account, which results in an exploring agent that actively queries information of the scene. In the second scenario, we add the instrumental term by which the agent makes an exploration-exploitation trade-off to reach the goal state as fast as possible.

4.1 Exploring behaviour

First, we focus on exploratory or information-seeking behaviour, i.e. actions are chosen based on the minimization of only the epistemic term of the expected free energy. For evaluation we restrict the robot arm to a fixed number of poses at a fixed height close to the table, so it can only observe a limited area of the workspace. The ground truth observations corresponding to the candidate poses are shown in a grid in Figure 2.

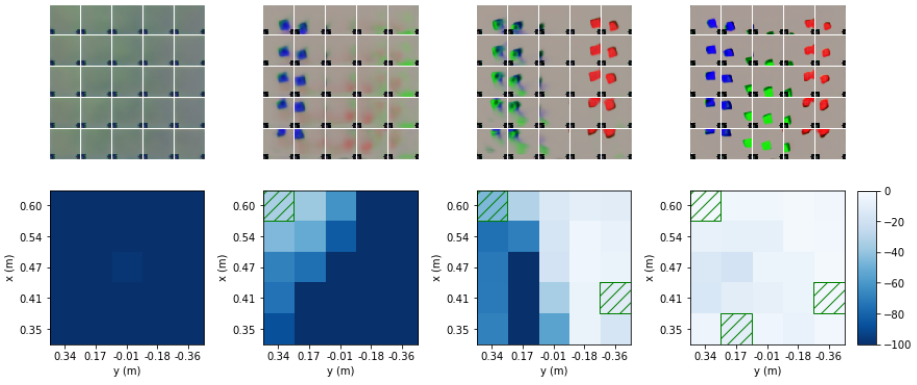
Initially, the agent has no information about the scene, and the initial state distribution $Q(\mathbf{s})$ is a zero-mean isotropic Gaussian. The expected observation is computed over 125 samples and visualized in the top row of Figure 3a. Clearly, the agent does not know the position of any of the objects in the scene, resulting in a relatively low value of the epistemic term from Equation (2) for all candiate poses. This is plotted in the bottom row of Figure 3a. The agent selects the upper left pose as indicated by the green hatched square in Figure 3b. After observing the blue cube in the upper left corner, the epistemic value of the left poses drops, and the robot queries a pose at the right side of the workspace. Finally, the robot queries one of the central poses, and the epistemic value of all poses becomes relatively high, as new observations do not yield more information. Notice that at this point, the robot can also accurately reconstruct the correct cubes from any pose as shown in the top row of Figure 3d.

4.2 Goal seeking behaviour

In this experiment, we use the same scene and grid of candidate poses, but now we provide the robot with a preferred observation from the red cube, indicated by the red hatched square in the bottom row of Figures 4a through 4d.

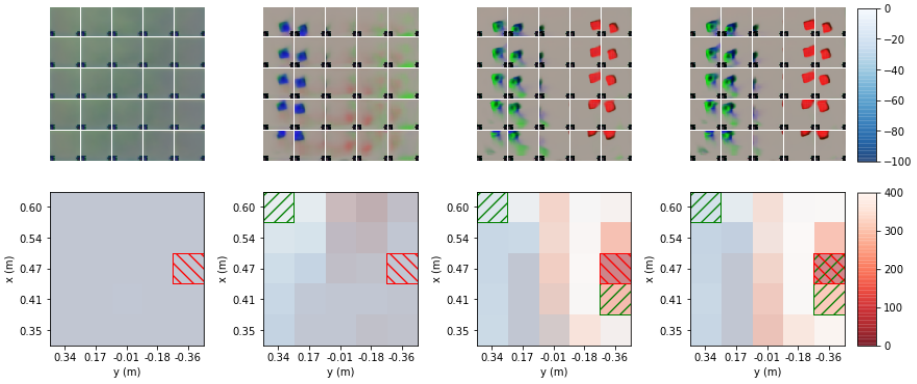
Initially, the agent has no information on the targets position and the same information-seeking behaviour from Section 4.1 can be observed in the first steps as the epistemic value takes the upper hand. However, after the second step, the agent has observed the red cube and knows which pose will reach the preferred state. The instrumental value takes the upper hand as indicated by the red values in Figures 4a through 4d. This is reflected by a significantly lower expected free energy. Even though the agent has not yet observed the green cube and is unable to create correct reconstructions as shown in Figure 4d, it will drive itself towards the preferred state. The trade off between exploratory and goal seeking behaviour can clearly be observed. In Figure 4c, the agent still has low epistemic values for the candidate poses to the left, but they do not outweigh the low instrumental value to reach the preferred state. The middle column of potential observations has a lower instrumental value, which is the result of using independent Gaussians for estimating likelihood on each pixel.

The number of steps to reach the preferred state is computed on 30 different validation scenes not seen in training, where the preferred state is chosen randomly. On average, the agent needs 3.7 steps to reach its goal. This is clearly



(a) $t = 0$ (b) $t = 1$ (c) $t = 2$ (d) $t = 3$

Fig. 3: The top row represents the imagined observations, i.e. the observations generated by the generative model, for each of the considered potential poses at a given step, the bottom row represents the epistemic value for the corresponding poses. Darker values represent a larger influence of the epistemic value. The green hatched squares mark the observed poses.



(a) $t = 0$ (b) $t = 1$ (c) $t = 2$ (d) $t = 3$

Fig. 4: The top row shows the imagined observations for each of the considered potential poses at a given time step. The bottom row shows the expected free energy for the corresponding poses. Blue is used to represent the epistemic value, while red is used to represent the instrumental value. The values of both terms are shown in the legend. The green hatched squares mark the observed poses, while the red hatched square marks the preferred state.

more efficient than a systematic grid search which would take on average 12.5 steps.

5 Conclusion

This work shows promising results in using the active inference framework for active vision. The problem is tackled with a generative model learned unsupervised from pure pixel data. The proposed approach can be used for efficiently exploring and solving robotic grasping scenarios in complex environments where a lot of uncertainty is present, for example in cases with a limited field of view or with many occlusions.

We show that it is possible to use learned latent space models as generative models for active inference. We show that both exploring and goal-seeking behaviour surfaces when using active inference as an action-selection policy. We demonstrated our approach in a realistic robot simulator and plan to extend this to a real world setup as well.

Acknowledgments This research received funding from the Flemish Government under the "Onderzoeksprogramma Artificiële Intelligentie (AI) Vlaanderen" programme.

References

1. Aloimonos, J., Weiss, I., Bandyopadhyay, A.: Active vision. International Journal of Computer Vision **1**(4), 333–356 (Jan 1988). https://doi.org/10.1007/bf00133571, https://doi.org/10.1007/bf00133571
2. Denzler, Zobel, Niemann: Information theoretic focal length selection for real-time active 3d object tracking. ICCV (2003)
3. Doumanoglou, A., Kouskouridas, R., Malassiotis, S., Kim, T.K.: 6D Object Detection and Next-Best-View Prediction in the Crowd. CVPR (2016)
4. Eslami et. al.: Neural scene representation and rendering. Science (2018)
5. Finn, C., Levine, S.: Deep visual foresight for planning robot motion. In: 2017 IEEE International Conference on Robotics and Automation (ICRA). pp. 2786–2793 (2017)
6. Friston, K., FitzGerald, T., Rigoli, F., Schwartenbeck, P., O'Doherty, J., Pezzulo, G.: Active inference and learning. Neuroscience & Biobehavioral Reviews (2016)
7. Heins, R.C., Mirza, M.B., Parr, T., Friston, K., Kagan, I., Pooresmaeili, A.: Deep active inference and scene construction (Apr 2020). https://doi.org/10.1101/2020.04.14.041129
8. Hepp, B., Dey, D., Sinha, S.N., Kapoor, A., Joshi, N., Hilliges, O.: Learn-to-score: Efficient 3D scene exploration by predicting view utility. ECCV (2018)
9. Kaba, M.D., Uzunbas, M.G., Lim, S.N.: A reinforcement learning approach to the view planning problem. CVPR (2017)
10. Mendoza, M., Vasquez-Gomez, J.I., Taud, H., Sucar, L.E., Reta, C.: Supervised Learning of the Next-Best-View for 3D Object Reconstruction. Pattern Recognition Letters (2020)
11. Mirza, M.B., Adams, R.A., Mathys, C.D., Friston, K.J.: Scene construction, visual foraging, and active inference. Frontiers in Computational Neuroscience (2016)
12. Nair, A.V., Pong, V., Dalal, M., Bahl, S., Lin, S., Levine, S.: Visual reinforcement learning with imagined goals. In: Bengio, S., Wallach, H., Larochelle, H., Grauman, K., Cesa-Bianchi, N., Garnett, R. (eds.) Advances in Neural Information Processing Systems 31, pp. 9191–9200. Curran Associates, Inc. (2018), http://papers.nips.cc/paper/8132-visual-reinforcement-learning-with-imagined-goals.pdf
13. Rasouli, A., Lanillos, P., Cheng, G., Tsotsos, J.K.: Attention-based active visual search for mobile robots. Autonomous Robots **44**(2), 131–146 (Aug 2019). https://doi.org/10.1007/s10514-019-09882-z, https://doi.org/10.1007/s10514-019-09882-z
14. Sancaktar, C., van Gerven, M., Lanillos, P.: End-to-end pixel-based deep active inference for body perception and action (2019)
15. Yamauchi, B.: A frontier-based exploration for autonomous exploration. ICRA (1997)
16. Çatal, O., Verbelen, T., Nauta, J., Boom, C.D., Dhoedt, B.: Learning perception and planning with deep active inference. In: ICASSP. pp. 3952–3956 (2020)

Appendix A The generative model

The generative model, described in this paper, is approximated by a neural network that predicts a multivariate Gaussian distribution with a diagonal covariance matrix. We consider a neural network architecture from the family of the variational autoencoders (VAE) [3, 6] which is very similar to the Generative Query Network (GQN) [1]. In contrast to the traditional autoencoders, this model encodes multiple observations into a single latent distribution that describes the scene. Given a query viewpoint, new unseen views can be generated from the encoded scene description. A high level description of the architecture is shown in Figure 5.

We represent the camera pose as 3D point and the orientation as a quaternion as this representation does not suffer from Gimbal lock. The encoder encodes each observation in a latent distribution which we choose to model by a multivariate Gaussian of 32 dimensions with a diagonal covariance matrix. The latent distributions of all observations are combined into a distribution over the entire scene in a similar way as the update step from the Kalman filter [2]. No prediction step is necessary as the agent does not influence the envrionment. In the decoder, the input is a concatenated vector of both the scene representation and the query viewpoint. Intuitively, both are important as the viewpoint determines which area of the scene is observed and the representation determines which objects are visible at each position. Between the convolutional layers, the intermediate representation is transformed using a FiLM layer, conditioned on the input vector, this allows the model to learn which features are relevant at different stages of the decoding process.

A dataset of 250 scenes, each consisting of approximately 25 (image, viewpoint) pairs has been created in a simulator in order to train this model. To limit the complexity of this model, all observations consist of the same fixed downward orientation.

Table 1: Training implementation details.

Optimizer	Adam
Learning rate	0.0001
Batch size	10
Number of observations	3-10
Tolerance	75.0
λ_{max}	100.0
λ_{init}	20.0

The neural network is optimized using the Adam optimizer algorithm with parameters shown in Table 1. For each scene between 3 and 10 randomly picked observations are provided to the model, from which it is tasked to predict a new one. The model is trained end-to-end using the GECO algorithm [7] on the following loss function:

$$\mathcal{L}_\lambda = D_{KL}[Q(\tilde{\boldsymbol{s}}|\tilde{\boldsymbol{o}})||\mathcal{N}(\mathbf{0}, \boldsymbol{I})] + \lambda \cdot \mathcal{C}(\boldsymbol{o}, \hat{\boldsymbol{o}}) \tag{3}$$

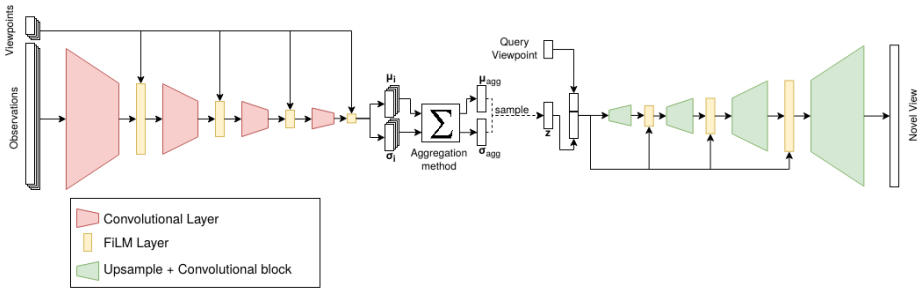


Fig. 5: Schematic view of the generative model. The left part is the encoder that produces a latent distribution for every observation, viewpoint pair. This encoder consists of 4 convolutional layers interleaved with FiLM [5] layers that condition on the viewpoint. This transforms the intermediate representation to encompass the spatial information from the viewpoint. The latent distributions are combined to form an aggregated distribution over the latent space. A sampled vector is concatenated with the query viewpoint from which the decoder generates a novel view. The decoder mimicks the encoder architecture and has 4 convolutional cubes (upsamples the image and processes it with two convolutional layers) interleaved with a FiLM layer that conditions on the concatenated information vector. Each layer is activated with a LeakyReLU [4] activation function.

The constraint $\mathcal{C}$ is applied to a MSE loss on the reconstructed and ground truth observation. This constraint simply means that the MSE should stay below a fixed tolerance. λ is a Lagrange multiplier and the loss is optimized using a min-max scheme [7]. Specific implementation values are shown in Table 1.

The expected free energy is computed for a set of potential poses. The generative model is first used to estimate the expected view for each considered pose. The expected value of the posterior with this expected view is computed for a large number of samples. This way, the expected epistemic term is computed. For numerical stability, we clamp the variances of the posterior distributions to a value of 0.25. The instrumental value is computed as the MSE between the preferred state and the expected observation. This essentially boils down to computing the log likelihood of every pixel is modelled by a Gaussian with a fixed variance of 1.

References

1. Eslami et. al.: Neural scene representation and rendering. Science (2018)
2. Kalman, R.E.: A new approach to linear filtering and prediction problems. Journal of Fluids Engineering, Transactions of the ASME **82**(1), 35–45 (1960). https://doi.org/10.1115/1.3662552

3. Kingma, D.P., Welling, M.: Auto-encoding variational bayes. 2nd International Conference on Learning Representations, ICLR 2014 - Conference Track Proceedings (Ml), 1–14 (2014)
4. Maas, A.L., Hannun, A.Y., Ng, A.Y.: Rectifier nonlinearities improve neural network acoustic models. In: in ICML Workshop on Deep Learning for Audio, Speech and Language Processing (2013)
5. Perez, E., Strub, F., De Vries, H., Dumoulin, V., Courville, A.: FiLM: Visual reasoning with a general conditioning layer. 32nd AAAI Conference on Artificial Intelligence, AAAI 2018 pp. 3942–3951 (2018)
6. Rezende, D.J., Mohamed, S., Wierstra, D.: Stochastic backpropagation and approximate inference in deep generative models. 31st International Conference on Machine Learning, ICML 2014 **4**, 3057–3070 (2014)
7. Rezende, D.J., Viola, F.: Taming vaes. CoRR **abs/1810.00597** (2018), http://arxiv.org/abs/1810.00597