Visual search as active inference

Emmanuel Daucé, Laurent U Perrinet

Visual search as active inference

Emmanuel Daucé[1,2][0000−0001−6596−8168] and Laurent Perrinet[1][0000−0002−9536−010X]

[1] Institut de Neurosciences de la Timone, CNRS/Aix-Marseille Univ, France.
[2] Ecole Centrale Marseille, France

Abstract. Visual search is an essential cognitive ability, offering a prototypical control problem to be addressed with Active Inference. Under a Naive Bayes assumption, the maximization of the information gain objective is consistent with the separation of the visual sensory flow in two independent pathways, namely the "What" and the "Where" pathways. On the "What" side, the processing of the central part of the visual field (the fovea) provides the current interpretation of the scene, here the category of the target. On the "Where" side, the processing of the full visual field (at lower resolution) is expected to provide hints about future central foveal processing given the potential realization of saccadic movements. A map of the classification accuracies, as obtained by such counterfactual saccades, defines a utility function on the motor space, whose maximal argument prescribes the next saccade. The comparison of the foveal and the peripheral predictions finally forms an estimate of the future information gain, providing a simple and resource-efficient way to implement information gain seeking policies in active vision. This dual-pathway information processing framework is found efficient on a synthetic visual search task with a variable (eccentricity-dependent) precision. More importantly, it is expected to draw connections toward a more general actor-critic principle in action selection, with the accuracy of the central processing taking the role of a value (or intrinsic reward) of the previous saccade.

Keywords: Object detection · Active Inference · Visual search · Visuomotor control · Deep Learning.

1 Introduction

Moving fast the eye toward relevant regions of the scene interestingly combines elements of action selection (moving the eye) with visual information processing. Noteworthy, the visual sensors have evolved during natural selection toward maximizing their efficiency under strong energy constraints. Vision in most mammals, for instance, has evolved toward a foveated sensor, maintaining a high density of photoreceptors at the center of the visual field, and a much lower density at the periphery. This limited bandwidth transmission is combined with a high mobility of the eye, that allows to displace the center of sight toward different parts of the visual scene, at up to 900 degrees per second in humans. Beyond

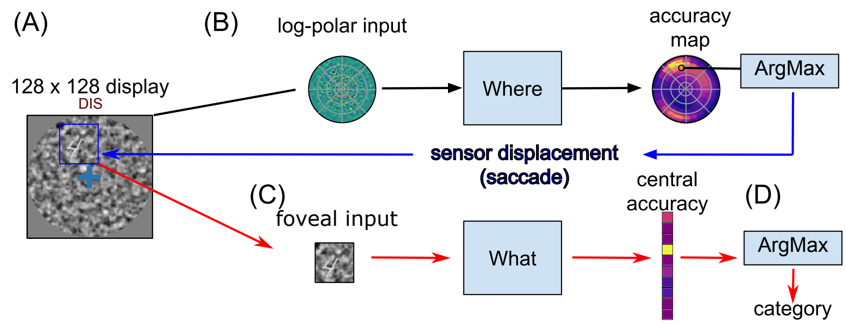

Fig. 1. Computational graph. Based on the anatomy of mammals' visual pathways, we define the following stream of information to implement visual search, one stream for localizing the object in visual space ("Where?"), the other for identifying it ("What?"). **(A)** The visual display is a stack of three layers: first a natural-like background noise is generated, characterized by noise contrast, mean spatial frequency and bandwidth [1]. Then, a sample digit is selected from the MNIST dataset [2], rectified, multiplied by a contrast factor and overlaid at a random position. Last, a circular, gray mask is put on. **(B)** The visual display is then transformed in a retinal input which is fed to the "Where" pathway. This observation is generated by a bank of filters whose centers are positioned on a log-polar grid and whose size increases proportionally with the eccentricity. The "Where" network outputs a collicular-like accuracy map. It is implemented by a three-layered neural network consisting of the retinal log-polar input, two hidden layers (fully-connected linear layers combined with a ReLU non-linearity) with 1000 units each. This map has a similar log-polar (retinotopic) organization and predicts the accuracy at the counter-factual positions of affordable saccades. The position of maximal activity in the "Where" pathway serves to generate a saccade denoted which displaces the center of gaze at a new position. **(C)** This generates a new sensory input in the fovea which is fed to a classification network ("What" pathway). This network is implemented using the three-layered LeNet neural network [2]. This network outputs a vector predicting the accuracy of detecting the correct digit. **(D)** Depending on the (binary) success of this categorical identification, we can first reinforce the What network, by supervisedly learning to associate the output with the ground truth through back-propagation. Then, we similarly train the "Where" network by updating its approximate prediction of the accuracy map.

the energetic efficiency, foveated vision improves the performance of agents by allowing them to focus on relevant vs. irrelevant information [3]. As such, this action perception loop uniquely specifies an AI problem [4, 5].

Indeed, Friston [6] proposed the FEP as a general explanatory principle behind the puzzling diversity of the mechanistic processes taking place in the brain and the body. One key ingredient to this process is the (internal) representation of counterfactual predictions, that is, the probable consequences of possible hypothesis as they would be realized into actions (here, saccades). Equipping the agent with the ability to actively sample the visual world allows to interpret saccades as optimal experiments, by which the agent seeks to confirm predictive models of the (hidden) world [4, 7]. Following such an active inference scheme, numerical simulations reproduce sequences of eye movements that fit well with empirical data [8, 9].

In particular, we focus here on *visual search* which is the cognitive ability to locate a single visual object in cluttered visual scene by placing the fovea on the object, in order to identify it [10–12]. As such, visual search intimately links the sampling of visual space (as it is done by the sensory apparatus) to the behavior which directs this sampling through the action of moving the direction of gaze. Note that the retina samples visual information predominantly on the fovea, though the target may lie in the periphery, where the acuity is lower. It is therefore commonplace that the target is not identifiable with the current information contained on the retinal image. As a consequence, visual search involves the problem that, given a limited observability, the object has to be localized *before* being identified.

Compared to earlier modelling studies, such as [13], we are concerned with the problem of both locating *and* identifying the target. This implies the capability to process the visual data and extract features from a complex (non-uniform) retinotopic visual sampling. This observation highlights an important hypothesis for solving the visual search problem. The semantic content of a visual scene is indeed defined by the positions and identities of the many objects that it contains. In all generality, the identity of an object is independent from its position in retinotopic space which is contingent on the observer's point of view. We thus consider the assumption that the visual system of mammals is built around such an independence hypothesis. The independence assumption, largely exploited in machine learning, is also known as the "Naïve Bayes" assumption. It simply considers as independent the different factors (or latent features) that explain the data. This implies here that inferring the identity and the position can be performed independently, and thus, could be processed *sequentially.* Selecting an object and identifying both its position and category may thus be the elementary bricks of visual processing. It may moreover explain the general separation of visual processing into the ventral and dorsal pathways. These two specific processing pathways are devoted to the processing of the stream of visual information, either to identify the semantic content of the visual field (the "What" pathway), or to decide where to orient next the line of sight (the "Where" pathway). They may operate in a continual and incremental turn-taking fashion, contributing to understand and exploit at best the visual information.

2 Problem statement: formalizing visual search as accuracy seeking

2.1 Visual search task

In this manuscript, we built upon an existing model [14] by precisely defining the mathematical framework under the Active Inference formalism. This model is based on a simplified generative model for a visual search task and a proposed algorithm to implement the task. First, in order to implement those principles into a concrete image processing task, we construct a simple yet ecological virtual experiment: After a fixation period of 200 ms, an observer is presented with a luminous 128×128 display showing a single target overlaid on a realistic noisy background (see Figure 1-A). This target is drawn in our case from the MNIST database of manuscript digits consisting of 60000 grayscale images of size 28×28 [2]. This image is displayed for a short period of about 500 ms which allows to perform (at most) one saccade toward the (unique) target. The goal of the agent is ultimately to correctly identify the digit.

2.2 Central processing

Following the Free Energy minimization principle (FEP) [6], engaging in a saccade stems on maintaining the visual field within the least surprising possible state. This implies, for instance, the capability to predict the next visual input through a generative model, and to orient the sight toward regions that minimize the agent's predicted model surprise [4]. Due to their limited memory and processing capabilities, living brains do not afford to predict or simulate their sensory environment exhaustively. Given the vast diversity of possible visual fields, one should assume that only the foveated part should deserve predictive coding. This implies that the saccadic motor control should be tightly optimized in order to provide a foveal data that should allow to accurately identify (and predict) the target.

In our model, we divide the retina into the fovea, which constitutes the center of the retina, and the peripheral region, which provides a visual information with a decreasing precision as a function of eccentricity. When considering the full visual field, the exponential decrease of the density of photo-receptors with respect to eccentricity [15] must be reflected in a non-uniform sampling of the visual data. It is here implemented as a log-polar conformal mapping, as it provides a good fit with observations in mammals and has a long history in computer vision and robotics [16]. These coordinates are denoted as the couple $u = (\epsilon, \theta)$ corresponding respectively to the log-eccentricity and azimuth in (spherical) polar coordinate by $\rho(u) \stackrel{\text{def.}}{=} (R \cdot \exp(\epsilon) \cdot \cos\theta, R \cdot \exp(\epsilon) \cdot \sin\theta)$ with R the maximal eccentricity.

Let us define as x a spatial coordinate in the input image cartesian referential (with $x_0 \stackrel{\text{def.}}{=} (0, 0)$ defining the center of the image), with x^t the position of the target and $k^t \in \{0, \ldots, 9\}$ its identity. The content of the fovea is considered as spatially uniform, here defined by extracting the 28×28 sub-image $f^t(x)$ at gaze direction x (initially x_0). At any given trial t drawn from the set $\mathcal{T}$ of trials of our virtual experiment, knowing the corresponding position x^t of the object, the problem of identifying the object can be solved, for instance, by a deep neural network [2] which infers its category. This network takes $f^t(x)$ as an input and returns a multinomial distribution vector $\mathbf{a}(f^t(x)) \in \mathbb{R}^{10}$ (with $\sum_k \mathbf{a}_k(f^t(x)) = 1$). This network takes here the role of the "What" pathway. Knowing the correct label k^t (and position x^t) for this trial t, this network is trained using a gradient descent with a categorical Cross Entropy loss. This loss is by definition:

$$\mathcal{L}_{\mathcal{K}}^t = -\log \mathbf{a}_{k^t}(f^t(x)) \tag{1}$$

The gradient descent is computed at each trial and the process is iterated over the set $\mathcal{T}$ of trials which give pairs of inputs $f^t(x)$ and outputs k^t for this supervised learning scheme. The accuracy of this classic neural network is known to exceed 98% over the genuine MNIST database [2], on par with human performance. Due to the max-pooling layers used between the convolutional layers, it also shows a robust translation invariance. In our experimental conditions, the network is trained over an augmented MNIST digits dataset, having a variable contrast, a variable shift (from 0 to 15 pixels away from the center) and a variable (randomly generated) background.

Knowing $f^t(x)$, the categorical response is $k^t = \arg\max_k \mathbf{a}_k(f^t(x))$. This response can be correct or incorrect. The correctness of the response is noted $\mathbf{o}(f^t(x))$ as we test our model. This value, that is 1 for a correct response and 0 otherwise, can be interpreted as a binary random variable. This random variable can be sampled at different t, with different success or failures depending on the actual target position x^t.

2.3 Accuracy map

The "What" neural network is constructed such that it can provide an estimate of the chance of success for every possible category by processing the central part of the visual field, i.e. the fovea. This chance of success could in principle be estimated the same way at any peripheral position $x \neq x_0$, through making a saccade and estimating the chance of success at gaze direction x. Then, for any target position x^t, and under an ergodic assumption, it could provide a belief on the average success that would be obtained at all positions x, i.e. $A^t(x) \overset{\text{def.}}{=} \mathbf{a}_{k^t}(f^t(x)) \approx Pr(k^t|f^t(x))$. This accuracy being defined for any gaze direction x, one could thus construct a *map* providing the expected probability of classification success knowing a potential future eye direction x afforded by a saccade. From the definition of the "What" network, this could be simply approximated by the accuracy of the selected class:

$$A^t(x) \approx \max_k \mathbf{a}_k(f^t(x)) \tag{2}$$

In principle, one could extract all possible sub-images $f^t(x)$ at all positions x, and estimate $A^t(x)$ directly. Moving the eye toward $\hat{x} = \arg\max_x A^t(x)$ and finding the object's identity at location $\hat{x}$ would solve the problem of both identifying and locating the target. This brute-force solution is of course computationally prohibitive, but provides a baseline toward a more biologically-relevant processing.

The belief in the success or the failure of identifying the target at different positions being, by construction, an output of the "What" pathway, it is essential for a visual search task to estimate the correctness of the test *prior* to a saccade, that is, to predict the statistics of $\mathbf{o}(f^t(x))$ from $A^t(x)$. The eye next position x being the result of a motor displacement u, with $x = x_0 + \rho(u) = \rho(u)$, it should be governed by a *policy*, i.e. a method that selects the next movement from the available visual input. The set of all possible displacements forms a *motor map*, and such a policy can be formalized as a mapping from the visual input space toward the motor map. Following the classical reinforcement learning literature, the motor map is expected to provide a value over the space of actions. We postulate here that *the value of the motor displacement u is identified with the classification accuracy obtained at position $x = \rho(u)$*. Moreover, we will show that, with minimal simplifying assumptions, this postulate can be framed into the more general framework of Active Inference.

3 Principles: supervised learning of action selection

3.1 Peripheral visual processing

On the visual side, local visual features are extracted as oriented edges as a combination of the retinotopic transform with filters resembling that found in the primary visual cortex [17]. The centers of these filters are radially organized around the center of fixation, with small receptive fields at the center and more large and scarce receptive fields at the periphery. The size of the filters increases proportionally with the eccentricity. To cover the visual space from the periphery to the fovea, we used 10 spatial eccentricity scales $\epsilon \in [-4, -1]$ such that the filters are placed at about 2, 3, 4.5, 6.5, 9, 13, 18, 26, 36.5, and 51.3 pixels from the center of gaze. There are 24 different azimuth angles allowing them to cover most of the original 128×128 image. At each of these positions, 6 different edge orientations and 2 different phases (symmetric and anti-symmetric) are computed.

This finally implements a bank of linear filters which models the receptive fields of the primary visual cortex. Assuming this log-polar arrangement, the resulting retinal visual data at this trial is noted as the feature vector $\mathbf{s}^t(x)$. For simplicity, it is noted $\mathbf{s}^t$ further on. The length of this vector is 2880, such that this retinal processing compresses the original image by about 83%, with high spatial frequencies preserved at the center and only low spatial frequencies conserved at the periphery. In practice, the bank of filters is pre-computed and placed into a matrix for a rapid transformation of input batches into feature vectors.

3.2 Motor control

Assuming the motor control is independent from the identity pathway, we take the classification success, as measured at the output of the "What" pathway, as the principal outcome of the "Where" pathway. It is assumed, in short, that the surprise should be higher in case of failure than in case of success, and that minimizing the surprise through active inference should be consistent with maximizing the likelihood of success.

On the motor side, a possible saccade location is defined as $u \overset{\text{def.}}{=} (\epsilon, \theta)$. Each coordinate of the visual field, except for the center, is mapped on a saccadic motor map. The motor map is also organized radially in a log-polar fashion, making the control more precise at the center and coarser at the periphery. This modeling choice is reminiscent of the approximate log-polar organization of the superior colliculus (SC) motor map [18]. Given a saccade command u, the corresponding classification success is noted $\mathbf{o}(f^t(\rho(u)))$. This success (or failure) being measured after the saccade, it must be guessed from a model. We posit here that the principle underlying the "Where" processing pathway is to predict the probability of success for every possible saccade command. This success is considered a realization of the likelihood $p(\mathbf{o}|u, \mathbf{s}^t)$. It is important here to note the dependence on the (peripheral) visual observation $\mathbf{s}^t$. Our likelihood function p can be seen as a mapping from $\mathbf{s}^t$ to the set $\mathcal{U}$ of possible saccade commands. Following these definitions, the objective of the "Where" processing pathway is to allow a saccadic decision by training such a likelihood function $w(u|\mathbf{s}^t)$ from observing failures and success from different saccades selection.

Now, the optimization being done on u, our saccade selection process relies on maximizing the likelihood of success, i.e. $\arg\max_u p(\mathbf{o} = 1|u, \mathbf{s}^t)$, that is consistent with assuming that a prior is put on observing a success, whatever the saccade. Computing a good approximation of the likelihood $p(\mathbf{o} = 1|u, \mathbf{s}^t)$ is therefore crucial to perform visual search:

$$w(u|\mathbf{s}^t) \approx p(\mathbf{o} = 1|u, \mathbf{s}^t) = A^t(\rho(u)) \tag{3}$$

where $\rho(u)$ is the future position of gaze for a saccade u, and $\mathbf{s}^t$ is the feature vector representing the present peripheral observation. The model predicts the accuracy of the "What" pathway, given the action u (saccade).

The choice of a saccade given the likelihood may be obtained from the maximum a posteriori rule :

$$\pi_{\max}(\mathbf{s}^t) = \arg\max_u p(\mathbf{o} = 1|u, \mathbf{s}^t) \cdot \Pr(u) \tag{4}$$

With for instance $\Pr(u) = \text{Unif}(u)$ a uniform prior probability on saccade selection, that is, uniformly on motor space, we obtain the policy (approximate in probability):

$$\pi_{\max}(\mathbf{s}^t) \approx \hat{\pi}_{\max}(\mathbf{s}^t) \stackrel{\text{def.}}{=} \arg\max_u w(u|\mathbf{s}^t) \tag{5}$$

Similarly, another strategy would be to use the approximate conditional expectation on action space:

$$\hat{\pi}_{\text{avg}}(\mathbf{s}^t) \stackrel{\text{def.}}{=} \int_u u \cdot w(u|\mathbf{s}^t) \cdot \Pr(u) \cdot du \tag{6}$$

Note that this conditional expectation is different from that that would operate in cartesian coordinates. In particular, using a log-polar accuracy map comes with an intrinsic prior for the saccades to be closer to the fixation point (see Figure 2).

Incidentally, the unimodal shape of the accuracy map indicates that a highest chance of success is found when the target is centered on the fovea, and for that reason the active inference mechanism should privilege saccades that will place the visual target at the center of the fovea. This is equivalent to identifying the location of the target in the retinotopic space, and thus inferring the spatial information from the visual field, with the future saccade taking the role of a latent variable explaining the current visual field $\mathbf{s}^t$.

From the active inference perspective, choosing the accuracy map as a likelihood function is like putting a prior on observing a success. In other words, the agent is more "surprised" in case of classification failure than in case of classification success. Taking the classification success as the principal outcome of the "Where" pathway, the action selection process now relies on minimizing the surprise as upper-bounded by the free-energy:

$$-\log p(\mathbf{o}=1|\mathbf{s}^t) \leq F \text{ with } F \stackrel{\text{def.}}{=} \mathbb{E}_q[-\log p(\mathbf{o}=1|\pi,\mathbf{s}^t)+\log q(\pi|\mathbf{s}^t,\mathbf{o}=1)-\log p(\pi|\mathbf{s}^t)] \quad (7)$$

with π the policy taking the role of a latent variable predicting the (future) classification success, and q being a probability distribution function on action selection policy. Finally, the visual search problem can be summarized as optimizing the function q which would define a saccade selection policy from a maximum success evidence perspective.

3.3 Higher level inference: choosing the processing pathway

Inferring the target location and identity sums up in our case to select a saccade in order to infer the target category from the future visual field. It is likely, however, that a saccade may not provide the expected visual data, and that a corrective saccade may be needed to improve the visual recognition. More generally, choosing to move the eye or to issue a categorical response from the available data resorts to select one processing pathway over the other: either realize the saccade or guess the category from the current foveal data. In order to make this choice, one must guess whether the chance of success is higher in the present, given the current visual field, or in the future, after the next saccade.

This, again, can be expressed under the active inference setup. Let $p(\mathbf{o}|f(x_0))$ the probability of success when processing the foveal data, as provided by the "What" network. Under the policy π (provided by the "Where" network), the decision decomposes into a binary choice between issuing a saccade or not. This decision should rely on comparing $p(\mathbf{o}|f(\rho(\pi(\mathbf{s}))$ (the future accuracy) and $p(\mathbf{o}|f(x_0))$ (the current accuracy). The active inference comes down here to a binary choice between actuating a saccade or "actuating" (testing) the categorical response.

Interestingly, the log difference of the two probabilities

$$\log p(\mathbf{o}|f^t(\rho(\pi(\mathbf{s}^t)) - \log p(\mathbf{o}|f^t(x_0)) \sim \log A^t(\rho(\pi(\mathbf{s}^t))) - \log A^t(x_0) \quad (8)$$

can be seen as an estimator of the *information gain* provided by the saccade. Choosing to actuate a saccade is thus equivalent to maximising the information gain provided by the new visual data, consistently with the classic "Bayesian surprise" metric [19]. Expanding over purely phenomenological models, our model finally provides a biologically interpretation of the information gain metric as a high-level decision criterion, linked to the comparison of the output of the two principal visual processing pathways.

3.4 Learning the accuracy map

Neural Networks are known to be in theory universal value function approximators and in practice, we will use a network architecture, alike to that used for the "What" pathway. This will provide a sufficient argument for showing that it is possible to learn such a mapping, while leaving open the possibility that other architectures may be actually implemented in the brain. The parametric neural network consists of the input feature vector $\mathbf{s}^t$ (of dimension 2880), followed by two fully-connected hidden layers of size 1000 with rectified linear activation units (ReLUs). A final fully-connected output layer with a sigmoid nonlinearity ensures that the output is compatible with a likelihood function. In accordance with observations [18, 20], the same log-polar compression pattern is defined at the retinal input and at the motor output (see Figure 1).

To learn the mapping provided by the "Where" network, we use the BCE cost as the Kullback-Leibler divergence between the tested accuracy and its approximation:

$$\mathcal{L}_{\mathcal{S}}^t = -[\mathbf{o}(f^t(\rho(u^t))) \cdot \log w(u^t|\mathbf{s}^t) + (1 - \mathbf{o}(f^t(\rho(u^t))) \cdot \log(1 - w(u^t|\mathbf{s}^t))] \quad (9)$$

We then optimize the parameters of the neural network implementing the "Where" pathway such as to optimize the approximation of the likelihood function. This can be achieved in our feed-forward model using back-propagation [2] with the input-output pairs $(\mathbf{s}^t, u^t)$ and the classification result as it is given by the "What" pathway. The role of the "What" pathway is here that of a critic of the output of the "Where" pathway (which takes the role of the actor). This separation of visuo-spatial processing into an actor and a critic is reminiscent of a more general actor-critic organization of motor learning in the brain, as postulated by Joel, Niv, and Ruppin [21].

The natural way to collect such supervision data is to draw data one by one in our virtual experiment, iteratively generating a saccade and computing the success of the detection. This is what would be performed by an agent which would sequentially learn by trial-and-error, using the actual recognition accuracy (after the saccade) to grade the action selection and leading to a reinforcement scheme. For instance, we could use corrective saccades to compute (a posteriori) the probability of a correct localization. In a computer simulation however, this calculation is slow and not amenable. To accelerate the learning in our scheme defined by a synthetic generative model, there exists however a computational shortcut to obtain more supervision pairs. Indeed, the learning of the where pathway may be done after that of the what pathway. Such a computational shortcut is allowed by the independence of the categorical performance with position. Moreover, for each input image, we know the true position in extrinsic (x^t) and intrinsic $(u^t \overset{\text{def.}}{=} \rho^{-1}(x^t))$ and identify k^t of the target. As such, one can compute the average accuracy map over the dataset and optimize equivalently

$$\mathcal{L}_{\mathcal{S}}^t = -\sum_{u \in \mathcal{S}} [A_0(u - u^t) \cdot \log w(u|\mathbf{s}^t) + (1 - A_0(u - u^t)) \cdot \log(1 - w(u|\mathbf{s}^t)] \quad (10)$$

where $A_0(\Delta u)$ stands for the accuracy map with respect to the true position, that is, of the accuracy when the input image to the "What" pathway is systematically shifted by $\rho(\Delta u)$. In our setting, this function varied little for different identities and we averaged it over all possible identities. Combining this translational shift and the shift-dependent accuracy map of the "What" classifier, the actual accuracy map at each trial can be thus predicted under an ergodic assumption by shifting the central accuracy map on the true position of the target (that is with $\Delta x = \rho(u) - x^t$). Then, this full accuracy map is a probability distribution function which can be computed on the rectangular grid of the visual display. We project this distribution on a log-polar grid to provide the expected accuracy of each hypothetical saccade in a retinotopic space similar to a collicular map. Applied to the full sized ground truth accuracy map computed in metric space, this gives an accuracy map at the different positions of the retinotopic motor space $\mathcal{S}$. This accelerate learning as it scales up both the set of tested saccade positions and gives the analog bias value instead of the binary outcome of the detection. Future work should explore if similar results will still hold when both networks are learned at the same time and with a trial-and-error strategy.

4 Results

After training, we observed that the "Where" pathway can correctly predict an accuracy map, whose maximal argument can be chosen to drive the eye toward a new viewpoint with a single saccade. There, a central snippet is extracted, that is processed through the "What" pathway, allowing to predict the digit's label. The full scripts for reproducing the figures and explore the results to the full range of parameters is available at https://github.com/laurentperrinet/WhereIsMyMNIST (under a GPLv3 license). The network is trained on 60 epochs of 60000 samples, with a learning rate equal to 10^{-4} and the Adam optimizer [22] with standard momentum parameters. An improvement in convergence speed was obtained by using batch normalization. One full training takes about 1 hour on a laptop. The code is written in Python (version 3.7.6) with the pyTorch library [23] (version 1.1.0).

Saccades distributions and classification success statistics resulting from this simple sequence are presented in Figure 2. Figure 2A-C provides an example of our active visual processing setup. The initial visual field (Fig. 2A) is processed through the "Where" pathway, providing a predicted accuracy map (compared with the true accuracy map in Fig. 2B)). The maximal argument of the accuracy map allows to actuate a saccade. The resulting visual field is provided in Fig. 2C, and the classification is done on the central part of the visual field only (red square). To generalize results, 1000 saccades are sampled for different sequences of input visual fields containing a target with a fixed eccentricity, but a variable identity, a variable azimuth and a variable background clutter. The digit contrast parameter is set to 70% and the eccentricity varies between 4 and 40 pixels. The empirical classification accuracies are provided in Figure 2D, for different eccentricities. These are averaged over all trials both on the initial central snippet and the final central snippet (that is, at the landing of the saccade).

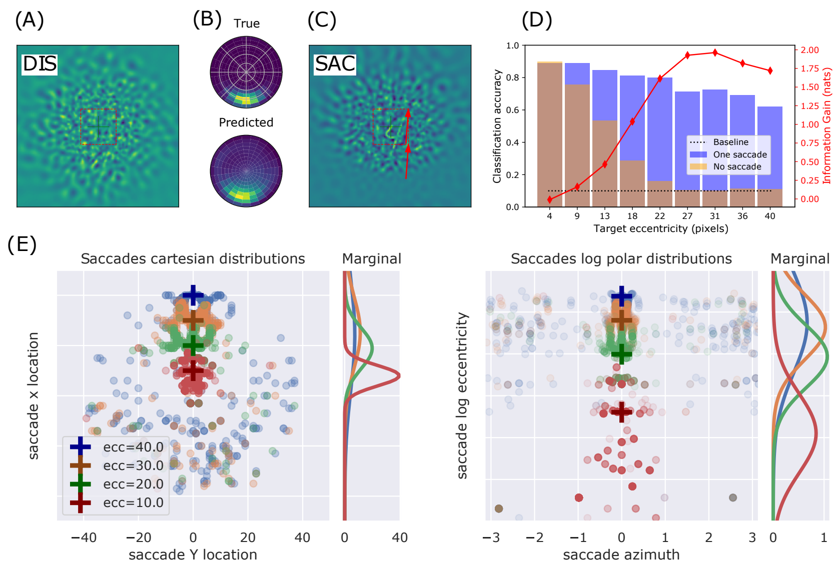

Fig. 2. Example of active vision after training the "Where" network. Digit contrast set to 70%. From left to right: **(A)** Magnified reconstruction of the visual input, as reconstructed from the primary visual feature vector through an inverse log-polar transform. **(B)** Color-coded radial representation of the output accuracy maps, with dark blue for the lower accuracy values, and yellow for higher values. The network output ("Predicted") is visually compared with the ground truth ("True"). **(C)** Visual field shift obtained after doing a saccade: The digit (the number 4) can now be recognized within the foveal region. **(D)** The final classification rate is plotted in function of the target eccentricity. The transparent orange corresponds to the pre-saccadic accuracy from the central classifier ('no saccade'). The blue bars correspond to the post-saccadic accuracy ('one saccade'), averaged over 1000 trials per eccentricity. Red line : empirical information gain, estimated from the accuracy difference. **(E)** Saccades distribution for different target eccentricities. The same saccades are plotted in (pixel) Cartesian coordinates on the left, and in log-polar coordinates on the right. The Cartesian coordinates correspond to the effector space while the log-polar coordinates correspond to the motor control space. In both cases, the empirical marginal distributions over one axis are shown on the right side.

The (transparent) orange bars provide the initial classification rate (without saccade) and the blue bars provide the final classification rate (after saccade). As expected, the accuracy decreases in both cases with the eccentricity, for the targets become less and less visible in the periphery. The decrease is rapid in the pre-saccadic case: the accuracy drops to the baseline level for a target distance of approximately 20 pixels from the center of gaze, consistent with the size of the target. The post-saccadic accuracy provides a much wider recognition range, with a slow decrease from about 90% recognition rate up to up to about 60% recognition when the target is put at 40 pixels away from the center. An estimate of the information gain provided is provided through a direct comparison of the empirical accuracies (red line). Here an optimal information gain is obtained in the 25-35 eccentricity range.

The lower accuracy observed at larger ranges is an effect of the visual signal bandwidth reduction at the larger eccentricities, that do not allow to accurately separate the target from the background. The spatial spreading of the saccades obtained at different eccentricities is represented on Figure 2E. The same saccades have been represented in Cartesian (pixel) coordinates (left figure) and in log-polar coordinates (right figure). By construction, the log-polar processing, implemented in the "Where" visuo-spatial pathway, leads to a decrease in saccade precision with respect to the eccentricity. This decreasing precision is illustrated by the higher variance of the saccades distribution observed at higher eccentricities, in the Cartesian space of the saccade realization. Interestingly, the variance of the marginal distribution of the saccades along the eccentricity axis is close to constant when represented in the log-polar space, that is, in the space of the (collicular) motor command. From 10 to 30 pixels away from the center, the precision of the command is invariant with respect to the eccentricity. The lower precision observed at about 40 pixels eccentricity only reflects a lower detection rate. Due to the log-polar construction of the motor map, the motor command (falsely) appears to display the same precision at various eccentricities. As it would be the case with a more detailed model of the motor noise, this log-polar organization of the control space can be interpreted as a natural renormalization, helping to counteract the precision loading that would otherwise be attached with the larger saccades, helping to provide a more uniform spread of the motor command in the effector space.

5 Discussion and perspectives

We proposed a computer-based framework allowing to implement visual search under bio-realistic constraints, using a foveated retina and a log-polar visuo-motor control map. A simple "Naïve Bayes" assumption justifies the separation of the processing in two pathway, the "What" visuo-semantic pathway and the "Where" visuo-spatial pathway. The predicted classification rate (or classification accuracy), serves as a guiding principle throughout the paper. It provides a way to link and compare the output of both pathways, serving either to select a saccade, in order to improve the chance of success, or to test a categorical response on the current visual data.

Future work should explore the application of this architecture to more complex tasks, and in particular to a more ecological virtual experiment consisting in classifying natural images. In particular, it would be possible to generalize this to a sequence of saccades, that is, mapping out an entire sequence of saccades by the where pathway, given the current field of view [24]. Finally, we used here the log-polar retinotopic mapping as a constraint originating from the anatomy of the visual pathways and have shown in Figure 2 that this implicitly generate a uniform action selection probabilty. At the temporal scale of natural selection, one could also consider this mapping as the emergence of an optimal solution considering an ecological niche, explaining for instance why foveal regions are more concentrated in predators than in preys, as shown for instance in avians [25]. As can be observed in the comparative study of pupils' shapes [26], this may justify the differences observed between preys (with a less sparse cone density at the periphery) and predators (with a tendency toward denser foveal regions) as a form. The compromise between the urgency to detect and the need to be accurate may justify the different balances which may exist in different species and thus as long term form of homeostasis [27].

References

1. Sanz-Leon, P., Vanzetta, I., Masson, G., and Perrinet, L.U.: Motion Clouds: Model-Based Stimulus Synthesis of Natural-like Random Textures for the Study of Motion Perception. Journal of Neurophysiology 107(11), 3217–3226 (2012). DOI: 10.1152/jn.00737.2011
2. Lecun, Y., Bottou, L., Bengio, Y., and Haffner, P.: Gradient-based learning applied to document recognition. Proceedings of the IEEE 86(11), 2278–2324 (1998). DOI: 10/d89c25
3. Tang, Y., Nguyen, D., and Ha, D.: Neuroevolution of Self-Interpretable Agents. Proceedings of the 2020 Genetic and Evolutionary Computation Conference (2020). DOI: 10/gg64b3
4. Friston, K.J., Adams, R.A., Perrinet, L.U., and Breakspear, M.: Perceptions as hypotheses: Saccades as experiments. Frontiers in Psychology 3 (2012). DOI: 10.3389/fpsyg.2012.00151
5. Daucé, E.: Active fovea-based vision through computationally-effective model-based prediction. Frontiers in Neurorobotics 12 (2018). DOI: 10/gfrhgj
6. Friston, K.: The free-energy principle: a unified brain theory? Nature reviews neuroscience 11(2) (2010)
7. Perrinet, L.U., Adams, R.A., and Friston, K.J.: Active inference, eye movements and oculomotor delays. Biological Cybernetics 108(6), 777–801 (2014). DOI: 10.1007/s00422-014-0620-8
8. Mirza, M.B., Adams, R.A., Mathys, C., and Friston, K.J.: Human visual exploration reduces uncertainty about the sensed world. PLOS ONE 13(1), e0190429 (2018). DOI: 10.1371/journal.pone.0190429
9. Cullen, M., Monney, J., Mirza, M.B., and Moran, R.: A Meta-Bayesian Model of Intentional Visual Search. (2020)
10. Treisman, A.M., and Gelade, G.: A Feature-Integration Theory of Attention. Cognitive psychology 12(1), 97–136 (1980). DOI: 10.1016/0010-0285(80)90005-5. 11957
11. Eckstein, M.P.: Visual search: A Retrospective. Journal of Vision 11(5), 14–14 (2011). DOI: 10/fx9zd9
12. Wolfe, J.M.: Visual Search. In: The Handbook of Attention. Pp. 27–56. MIT Press, Cambridge, MA, US(2015)

13. Najemnik, J., and Geisler, W.S.: Optimal Eye Movement Strategies in Visual Search. Nature 434(7031), 387–391 (2005). DOI: 10/bcbw2b
14. Daucé, E., Albiges, P., and Perrinet, L.U.: A dual foveal-peripheral visual processing model implements efficient saccade selection. Journal of Vision 20(8), 22–22 (2020). DOI: 10.1167/jov.20.8.22
15. Watson, A.B.: A formula for human retinal ganglion cell receptive field density as a function of visual field location. Journal of vision 14(7), 15–15 (2014)
16. Javier Traver, V., and Bernardino, A.: A review of log-polar imaging for visual perception in robotics. Robotics and Autonomous Systems 58(4), 378–398 (2010)
17. Fischer, S., Sroubek, F., Perrinet, L.U., Redondo, R., and Cristóbal, G.: Self-invertible 2D log-Gabor wavelets. International Journal of Computer Vision 75(2), 231–246 (2007). DOI: 10.1007/s11263-006-0026-8
18. Sparks, D.L., and Nelson, I.S.: Sensory and motor maps in the mammalian superior colliculus. Trends in Neurosciences 10(8), 312–317 (1987)
19. Itti, L., and Baldi, P.: Bayesian surprise attracts human attention. Vision Research 49(10), 1295–1306 (2009). DOI: 10.1016/j.visres.2008.09.007
20. Connolly, M., and Van Essen, D.: The representation of the visual field in parvocellular and magnocellular layers of the lateral geniculate nucleus in the macaque monkey. Journal of Comparative Neurology 226(4), 544–564 (1984)
21. Joel, D., Niv, Y., and Ruppin, E.: Actor–critic models of the basal ganglia: New anatomical and computational perspectives. Neural networks 15(4-6), 535–547 (2002)
22. Kingma, D.P., and Ba, J.: Adam: A method for stochastic optimization. arXiv preprint arXiv:1412.6980 (2014)
23. Paszke, A., Gross, S., Massa, F., Lerer, A., Bradbury, J., Chanan, G., Killeen, T., Lin, Z., Gimelshein, N., Antiga, L., Desmaison, A., Kopf, A., Yang, E., DeVito, Z., Raison, M., Tejani, A., Chilamkurthy, S., Steiner, B., Fang, L., Bai, J., and Chintala, S.: PyTorch: An Imperative Style, High-Performance Deep Learning Library. In: Advances in Neural Information Processing Systems 32. Ed. by H. Wallach, H. Larochelle, A. Beygelzimer, F. dAlché-Buc, E. Fox, and R. Garnett, pp. 8024–8035. Curran Associates, Inc.(2019)
24. Hoppe, D., and Rothkopf, C.A.: Multi-Step Planning of Eye Movements in Visual Search. Scientific Reports 9(1), 144 (2019). DOI: 10/gfwcvc
25. Moore, B.A., Tyrrell, L.P., Pita, D., Bininda-Emonds, O.R.P., and Fernández-Juricic, E.: Does Retinal Configuration Make the Head and Eyes of Foveate Birds Move? Sci Rep 7 (2017). DOI: 10/f9k78h
26. Banks, M.S., Sprague, W.W., Schmoll, J., Parnell, J.A.Q., and Love, G.D.: Why Do Animal Eyes Have Pupils of Different Shapes? Science Advances 1(7), e1500391 (2015). DOI: 10/gg66t6
27. Connant, R.C., and Ashby, W.R.: Every Good Regulator of a System Must Be a Model of That System. International Journal of Systems Science 1(2), 89–97 (1970). DOI: 10/bbgr9b