# Learning Where to Look for What to See : A Foveated Visual Search Model

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

[1] Institut de Neurosciences des Systèmes, CNRS/Aix-Marseille Université, France
[2] Institut de Neurosciences de la Timone, CNRS/Aix-Marseille Université, France

**Abstract.** In computer vision, the visual search task consists in extracting a scarce and specific visual information (the "target") from a large and crowded visual display. This task is usually implemented by scanning the different possible target identities at all possible spatial positions, hence with strong computational load. The human visual system employs a different strategy, combining a foveated sensor with the capacity to rapidly move the center of fixation using saccades. Saccade-based visual exploration can be idealized as an inference process, assuming that the target position and category are independently drawn from a common generative process. Knowing that process, visual processing is then separated in two specialized pathways, the "where" pathway mainly conveying information about target position in peripheral space, and the "what" pathway mainly conveying information about the category of the target. We consider here a dual neural network architecture learning independently where to look and then at what to see. This allows in particular to infer target position in retinotopic coordinates, independently to its category. This framework was tested on a simple task of finding digits in a large, cluttered image. Simulation results demonstrate the benefit of specifically learning where to look before actually knowing the target category. The approach is also energy-efficient as it includes the strong compression rate performed at the sensor level, by retina and V1 encoding, which is preserved up to the action selection level, highlighting the advantages of bio-mimetic strategies with regards to traditional computer vision when computing resources are at stake.

**Keywords:** Object localization · Active Inference · Visual search · Visuomotor control · Deep Learning

## 1 Introduction

*Problem statement.* The promise of artificial vision to identify objects in natural images is ever increasing. Image processing algorithms recently outreached the performance of human observers in specific image categorization tasks (He et al. 2015). Initially trained on energy greedy, high performance computers, they are now designed to work on more common hardware such as desktop computers with a decent GPU (Sandler et al. 2018). However, these algorithms are still far from human performances, even for simple tasks. Take for instance the case of an encounter with a friend in a crowded café. To catch the moment at which she arrives, you need to visually search for her specific face despite all the remaining sensory clutter. To do so, you need to scan relevant parts of the visual scene with your gaze. Doing a saccade at these locations will allow you to recognize your friend. The main difficulty of this task is to learn to categorize this particular object class given all possible spatial configurations and respective geometrical visual transformations.

This visual search experience can be formalized and simplified in a way reminiscent to classical psychophysical experiments: an observer is asked to classify digits

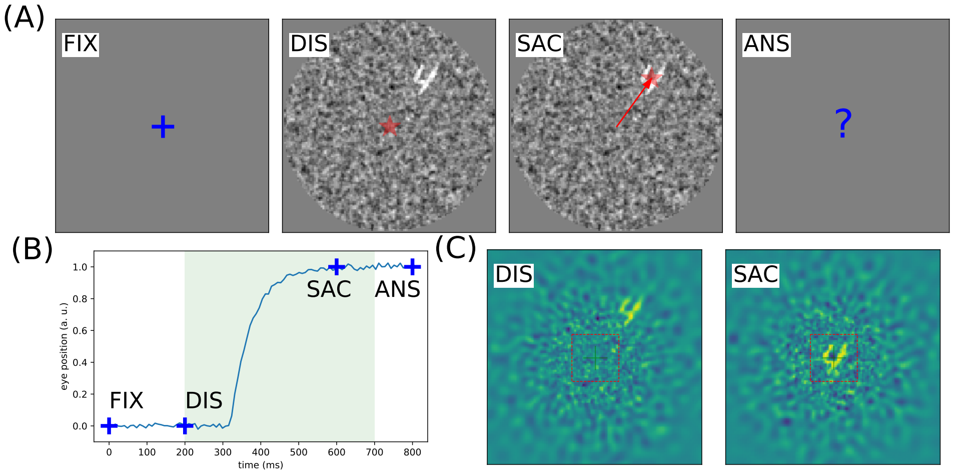

**Fig. 1. Problem setting**: In generic, ecological settings, the visual system faces a tricky problem when searching for one target (from a class of targets) in a cluttered environment. It is synthesized in the following experiment: **(A)** After a fixation period `FIX` of 200 ms, an observer is presented with a luminous display `DIS` showing a single target from a known class (here digits) and at a random position. The display is presented for a short period of 500 ms (light shaded area in B), that is enough to perform at most one saccade (here, successful) on the potential target `SAC`. Finally, the observer has to identify the digit by a keypress `ANS`. **(B)** Prototypical trace of a saccadic eye movement to the target position. In particular, we show the fixation window `FIX` and the temporal window during which a saccade is possible (green shaded area). **(C)** Simulated reconstruction of the visual information from the (interoceptive) retinotopic map at the onset of the display `DIS` and after a saccade `SAC`, the dashed red box indicating the visual area of the "what" pathway. In contrast to an exteroceptive representation (see A), this demonstrates that the position of the target has to be inferred from a degraded (sampled) image. In particular, the configuration of the display is such that by adding clutter and reducing the size of the digit, it may become necessary to perform a saccade to be able to identify the digit. The computational pathway mediating the action has to infer the location of the target *before seeing it*, that is, before being able to actually identify the target's category from a central fixation.

(for instance as taken from the MNIST database) as they are shown on a computer display. However, these digits can be placed at random positions on the display, and visual clutter is added as a background to the image (see Figure 1-A). This opens the possibility that the position of the object may be detected in the clutter without being identified in the first place (see Figure 1-C). This defines more precisely our problem: how do we localize an object in a large image while knowing *a priori* its category but not its identity? This generic visual search problem is of broad interest in machine learning, computer vision and robotics, but also in neuroscience, as it speaks to the mechanisms underlying foveation and more generally to low-level attention mechanisms.

Inherent to this problem is the combinatorial explosion implied by an increasing number of parameters. State-of-the art classification architectures consequently contain many millions parameters with subsequent energy consumption increase while still handling relatively small images. This introduces a trade-off between efficiency and average accuracy, for instance in autonomous driving such that the algorithm is fast enough to detect visual objects in a glance while running on resource-constrained devices like embedded devices. Globally, this performance is still lower than that of humans. Indeed, the human visual system can perform such a feat both rapidly, – in less than 100 ms (Kirchner and Thorpe 2006) – and at a low energy cost ($< 5\ W$). On top of that, it is mostly self-organized, robust to visual transforms or lighting conditions and can learn with a few examples. If many different anatomical features may explain this efficiency, a main difference lies in the fact that its sensor (the retina) combines a non homogeneous sampling of the world with the capacity to rapidly change its center of fixation. Indeed, on the one hand, the retina is composed of two separate systems: a central, high definition fovea (a disk of about 6 degrees of diameter in visual angle around the center of gaze) and a large, lower definition peripheral area. On the other hand, the retina is attached on the back of the eye which is capable of low latency, high speed eye movements. In particular, saccades allow for efficient changes of the position of the center of gaze: they take about 200 ms to initiate, last about 200 ms and usually reach a maximum velocity of approx 600 degrees per second. This behavior is prevalent during our lifetime (about a saccade every 2-3 seconds, that is, almost a billion saccade in a lifetime). The interplay of those two features allows human observers to engage in an integrated action perception loop which sequentially scans and analyses the different parts of the image.

*State of the art.* To take advantage of this visuomotor behavior, it is of particular importance to understand both its computational and neurophysiological principles. First, the joint problem of target localization and identification is a classical problem of visual search in computer vision. It is very general and may address apparently simple questions such as “find the green bottle on the table”. When restricted to a mere “feature search” (Treisman and Gelade 1980), many solutions are proposed. Notably, recent advances in deep-learning have provided efficient models such as faster-RCNN (Ren et al. 2017) or YOLO (Redmon et al. 2015). This last implementation is particularly interesting for our sake as it predicts in the image the probability of proposed bounding boxes around visual objects. While rapid, the number of boxes greatly increases with image size and necessitates dedicated hardware. In parallel, when limited to a few objects of interest in the image, this strategy amounts to a classical problem in neuroscience, that is, the transformation of a luminous image into a saliency map (Itti and Koch 2001), essential to understand and predict saccades, but also to serve as phenomenological models of attention. The saliency approach was recently extended using deep learning to estimate saliency maps over large databases of natural images (Kümmerer, Wallis, and Bethge 2016). While these methods are efficient at predicting the probability of fixation, they miss an essential component in the action perception loop: they operate on the full image while the retina operates on the non-uniform, foveated sampling of visual space (see Figure 1-B). Herein, we believe that this fact is an essential factor to reproduce and understand this active vision process.

In contrast to phenomenological (or “bottom-up”) approaches, models of active vision (Butko and Movellan 2010; Friston et al. 2012; Najemnik and Geisler 2005) provide the ground principles of saccadic exploration. They assume the existence of a generative model from which both the target position and category can be inferred through active sampling. This comes from the constraint that the visual sensor is foveated but can generate a saccade. Several studies are relevant to our endeavor. First, one can consider optimal strategies to solve the problem of the visual search of a target (Najemnik and Geisler 2005). In a setting similar to that presented in Figure 1-A, where the target is an oriented edge and the background is defined as pink noise, authors show first that a Bayesian ideal observer comes out with an optimal strategy, and second that human observers are close to that optimal performance. Though well predicting sequences of saccades in a perception action loop, this model is limited by the simplicity of the display (elementary edges added on stationary noise, a finite number of locations on a discrete grid) and by the abstract level of modeling. Despite these (inevitable) simplifications, this study could successfully predict some key characteristics of visual scanning such as the trade-off between memory content and rapidity. Looking more closely at neurophysiology, the study of (Samonds, Geisler, and Priebe 2018) allows to go further in understanding the interplay between saccadic behavior and the statistics of the input. In this study, authors were able to manipulate the size of the saccades by monitoring key properties of the presented (natural) images. For instance, smaller images generate smaller saccades. Interestingly, they also predicted the size of saccades for different species, including mice which lack a foveal region, from the size of visual receptive fields. One key prediction of this study which is relevant for our problem is the fact that saccades seem optimal to *a priori* decorrelate the visual input, that is, to minimize redundancy in the sequence of generated saccades, knowing the statistics of the visual inputs.

A further modeling perspective is provided by (Friston et al. 2012). In this setup, a full description of the visual world is used as a generative process. An agent is completely described by giving the generative model governing the dynamics of its internal beliefs and is interacting with this image by scanning it through a foveated sensor, just as described in Figure 1. Thus, 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. 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). Following such an active inference scheme (Mirza et al. 2018) numerical simulations reproduce sequential eye movements that fit well with empirical data. Saccades are here a consequence of an active seek for the agent to minimize the uncertainty about his beliefs, knowing his priors on the generative model of the visual world.

*Outline.* Stemming from the active vision general principles, our aim is to produce a principled model that may both explain the essential features of human vision and provide ways toward efficient computer implementations. We also aim at reunifying the fragmentation of the many different approaches respective to their fields (Machine learning, neuroscience, robotics), and envisage an integrated computational model of foveated active vision. It is known that inverting a generative model over a large (one-step ahead) hypothesis space of all possible saccades is computationally-intensive. In particular, complex combinatorial inferences are here replaced by separate pathways, i.e. the spatial ("where") and categorical ("what") pathways, whose knowledge is combined to infer optimal eye displacements and subsequent identification of the target. In addition, the agent is equipped with a foveated sensor, to which also contributes to minimizing the overall computational cost of finding a target. Taking such priors, we optimize the behavior of this agent and explore its key properties.

To that aim, this paper is organized as follows: After this introduction, we define the principles underlying accuracy-based saccadic control in section 2. We first define notations, variables and equations for the generative process governing the experiment and the generative model for the active vision agent. In particular, we derive our method to simplify the learning of an optimal agent given these definitions. In section 3, implementation details are given, providing ways to reproduce our results. In section 4, preliminary results of numerical simulations of the agent are presented, demonstrating the applicability of this framework to different task complexity levels. This allows us to derive some limits of the agent and, as in (Najemnik and Geisler 2005), we draw some analogies with biologically observed eye movements. Finally, in section 5, we summarize these results in comparison with

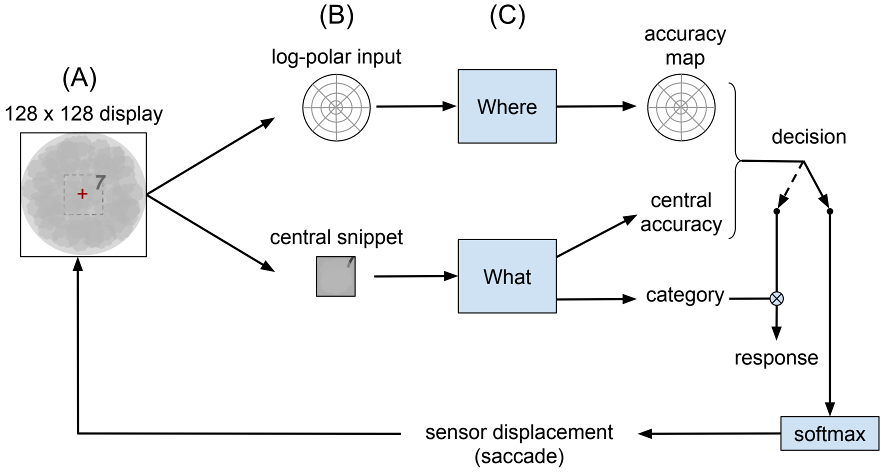

**Fig. 2. Methods for simulating active vision**: **(A)** We first define the model which generates images. It is composed of three different random processes: one choosing a sample image from the MNIST database (of size $28 \times 28$) and placing it at a random position within the circular mask on the $128 \times 128$ display. Then, this image is rectified and multiplied by a contrast factor and finally embedded in a natural-like noise characterized by noise contrast, mean spatial frequency and bandwidth (Sanz-Leon et al. 2012) (see an example in Figure 1-A, `DIS`). **(B)** The full-sized image is transformed into a retinal image which will be fed to the "where" pathway. This transform is implemented by a bank of filters whose centers are positioned on a log-polar grid and whose radius increases proportionally with eccentricity. In addition, a similar topographic map is used to represent the accuracy of each hypothetical position of a saccade, as represented by the collicular map. **(C)** The "where" pathway is implemented by a three-layered neural network consisting of the retinal input, two hidden layers with 1000 units each and a collicular output. Each unit is associated with a ReLU non-linearity. To learn to associate the output of the network with the ground truth, supervised training is performed using back-propagation with a binary cross entropy loss. This scalar measures the distance between both distributions (it is always positive and null if and only if they are equal). The position of maximal activity the "where" pathway serves to move the center of gaze and classify the foveal image using the "what" pathway.

other similar schemes. We conclude by showing the relative advantages of using this active inference approach.

## 2 Principles

In this study, the visual scene is made of a target object placed in the foreground at a random position over a noisy background. An agent controls a foveal visual sensor that can move over the visual scene through saccades (see Figure 1). The agent aims at understanding the visual scene, here identifying both the target position and identity from visual samples.

### 2.1 Active inference

Active inference assumes a hidden external state $e$, which is known indirectly through its effects on the sensor. The external state corresponds to the physical environment. The visual field $x$ is the state of the sensors, that is, a partial view of the visual scene, measured through a generative process : $x \sim p(X|e)$. The real physical state $e$ being hidden, a parametric model $\theta$ is assumed to allow for an estimate of the cause of the current visual field through model inversion thanks to Bayes formula, in short:

$$p(E|x) \propto p(x|E;\theta)$$

The external state is assumed to split in two (independent) components, namely $e = (u, y)$ with $u$ the interoceptive body posture (in our case the gaze orientation) and $y$ the object shape (or object identity). It is also assumed that a set of motor commands $A = \{..., a, ...\}$ (here saccades) may control the body posture, but not the object's identity, so that $y$ is invariant to $a$.

In a predictive setup, the consequence of every saccade should be analyzed through model inversion *over the future observations*, that is, predicting the effect of every action to choose the one that may optimize future inferences. The benefit of each action should be quantified through a certain metric (future accuracy, future posterior entropy, future variational free energy, ...), that depend on the current inference $p(U, Y|x)$. The saccade $a$ that is selected thus provides a new visual sample from the scene statistics. If well chosen, it should improve the understanding of the scene (here the target position and category). However, estimating in advance the effect of every action over the range of every possible object shapes and body postures is combinatorially hard, even in simple cases such as vision, and thus infeasible in practice.

The predictive setup necessitates in practice to restrain the generative model in order to reduce the range of possible combinations. One such restriction, known as the “Naïve Bayes” assumption, considers the independance of the factors that are the cause of the sensory view. The independence hypothesis allows considering the position $u$ and the category $y$ being independently inferred from the current visual field, i.e $p(U, Y|x) = p(U|x)p(Y|x)$. This property is strictly true in our setting and is very generic in vision for simple classes (such as digits) and simple displays (but see (Võ and Wolfe 2012) for more complex visual scene grammars). This independence assumption allows to separate the scene analysis in two independent tasks. A first task consists in identifying the target (namely inferring $y$ from $x$) and a second task consists in localizing the target (namely inferring $u$ from $x$). Each task is moreover assumed to be realized in parallel through distinct computational pathways, that are referred as the ‘What” and the “Where” pathways by analogy with the brain (see figure 2) However, we will here simplify the setting by considering only one possible saccade. Each pathway is here assumed to rely on different sensor morphologies. By analogy with biological vision, the target identification is assumed to rely on the very central part of the retina (the fovea), that comes with higher cones density, and thus higher spatial precision. In contrast, the target localization should rely on full visual field, with peripheral regions having a lesser sensor density and a lesser sensitivity to high spatial frequencies.

### 2.2 Metric training

Next, the effect of a saccade is to shift the visual field from one place to another. Concretely, each saccade provokes a new visual field $x'$ and a new subjective position $u'$, while the target identity $y$ remains unchanged. Examining the current visual field $x$ allows to form two hypotheses, namely $p(U|x)$ and $p(Y|x)$. It may happen, however, that the current inferences may not be accurate enough and there may be a “better” eye direction from which more evidence could be grabbed, i.e. it may be worth issuing a saccade so that $p(U'|x')$ and $p(Y|x')$ should be more accurate. Choosing the next saccade thus means using a model to predict how accurate $p(U|x)$ and $p(Y|x)$ will be after the saccade realization. It is worth noting that active inference needs either the current identity $y$ or the current eye direction $u$ to be readable from the present view, in order to effectively predict future inferences, through computationally intensive predictions. In detail, modeling the full sequence of operations that lead to both estimate $p(U'|x')$ and $p(Y|x')$ means predicting the future visual field $x'$ over all possible saccades, that may be tremendously costly in case of large visual fields. Better off instead is to form a statistics over the (scene understanding) benefit obtained from past saccades in the same context, that is forming an *accuracy map* from the current view. This is the essence of the *sampling-based metric prediction* that we develop here. The putative effect of every saccade should be condensed in a single number, the *accuracy*, that quantifies the final benefit of issuing saccade $a$ from the current observation $x$. If $a$ is a possible saccade and $x'$ the corresponding future visual field, the result of the categorical classifier over $x'$ can either be correct (1) or incorrect (0). If this experiment is repeated many times over many visual scenes, the probability of correctly classifying the future visual field $x'$ from $a$ forms a probability, i.e. a number between 0 and 1, that reflects the proportion of correct and incorrect classifications. It more or less corresponds to inferring the true target identity $\hat{y}$, i.e. $p(\hat{y}|x')$, including the update of the eye direction, that is a sample of the "real" generative process. In a biological setting, this would be acchieved for instance by catch-up saccades that would scan the area neighboring the saccade that was actually issued. To sum up, a main assumption here is that instead of trying to detect the actual position of the target, it is better for the agent to estimate how accurate the categorical classifier will be after moving the eye. This forms an accuracy map that may be learned through trials and errors, by actuating saccades and taking the final classification success or failure as a teaching signal. Such a *predictive accuracy map* is assumed to be the core of a realistic saccade-based vision system.

Note that compared to a brute force approach which would scan for all possible positions in an image, this map should compress the information, as exemplified by a retinotopic map. The map should be mostly organized radially, preserving the initial retinotopic organization. The operations that transform the initial primary visual data should preserve the initial retinotopic organization, so as to form a final retinotopic accuracy map (see figure 2). Accordingly with the initial data, the retinotopic accuracy map may thus provide more detailed accuracy predictions in the center, and coarser accuracy predictions in the periphery. Finally, each different initial visual field may bring out a different accuracy map, indirectly conveying information about the target retinotopic position. A final action selection (motor map) should then overlay the accuracy map through a winner-takes-all mechanism, implementing the saccade selection in biologically plausible way, as it is thought to be done in the superior colliculus, a brain region responsible for oculo-motor control [TODO:ref needed].

### 2.3 Information gain and the inhibition of return

From the information theory standpoint, each saccade comes with fresh visual information about the visual scene that can be quantified by an *information gain*, namely:

$$\mathrm{IG}_{\max} = \max_{u'} \log p(y|u', x', x, u) - \log p(y|x, u) \tag{1}$$

$$\simeq \max_{u'} \log p(y|x') - \log p(y|x) \tag{2}$$

with the left term representing the future accuracy (after the saccade is realized) and the right term representing the current accuracy as it is obtained from the 'what' pathway. The accuracy gain may be averaged over many saccades and many initial eccentricities (so that the information gain may be close to zero when the initial $u$ is very central).

For the saccade is subject to predictions errors and execution noise, the actual $u'$ may be different from the initial prediction. The final accuracy, as instantiated in the accuracy map, contains this intrinsic imprecision, and is thus necessary lower than the optimal one. The consequence is that in some cases, the approximate information gain may become negative, when the future accuracy is actually lower than the current one. This is for instance the case when the target is centered on the fovea. This should encourage the agent to select a saccade "away" from the central position, which is reminiscent of a well-known phenomenon in vision known as the "inhibition of return" (Itti and Koch 2001). Combining accuracy predictions from each pathway may thus allow to refine saccades selection in a way that complies with biological vision.

## 3 Implementation

To test the validity of our hypothesis, let us find a function implementing the "where" network. This function should be able to find the position of an object knowing only the degraded retinal image. Here, we describe the methods that we will follow to find that function, from the generative models (first external and then internal) to the actual implementation of the "where" pathway.

### 3.1 Exteroceptive Generative model

We define here the generative model for input display images as shown first in Figure 1-A (`DIS`) and as implemented in Figure 2-A.

*Targets.* Following a common hypothesis regarding active vision, visual scenes consist of a single visual object of interest. We use the MNIST database of handwritten digits introduced by (Lecun et al. 1998). Samples are drawn from a database of 60000 grayscale $28 \times 28$ pixels images and separated between a training and a validation set (see below the description of the "where" network).

*Full-scale images.* Each sample position is draw a random in a full-scale image of size $128 \times 128$. To enforce isotropic saccades, a centered circular mask covering the image (of radius 64 pixels) is defined, and the position is such that the embedded sample fits entirely into that circular mask.

*Background noise setting.* To implement a realistic background noise, we generate synthetic textures (Sanz-Leon et al. 2012) using a bi-dimensional random process. The texture is designed to fit well with the statistics of natural images. We chose an isotropic setting where textures are characterized by solely two parameters, one controlling the median spatial frequency $sf_0$ of the noise, the other controlling the bandwidth around the central frequency. Equivalently, this can be considered as the band-pass filtering of a random white noise image. Finally, these images are rectified to have a normalized contrast.

*Mixing the signal and the noise.* Finally, both the noise and the target image are merged into a single image. Two different strategies are used. A first strategy emulates a transparent association, with an average luminance computed at each pixel, while a second strategy emulates an opaque association, choosing for each pixel the maximal value. The quantitative difference was tested in simulations, but proved to have a marginal importance.

### 3.2 Interoceptive generative model

We now define the simplified anatomy of the agent, which is composed of two separate pathways.

*Foveal vision and the "what" pathway* First, foveal vision is defined as the 28 × 28 pixels image centered at the point of fixation (see dashed red box in Figure 1-C). This image is then directly passed to the agent's visual categorical pathway (the "What" pathway). This is realized by the known "LeNet" classifier (Lecun et al. 1998), that processes the $28 \times 28$ central pixels to identify the target category. Such a network is directly provided (and unmodified) by the pyTorch library (Paszke et al. 2017), and consists of a 3-layered Convolutional Neural Network. It is trained over the (centered) MNIST database after approx 20 training epochs. The network outputs a vector representing the probability of detecting each of the 10 digits. We use the argument of the output neuron with maximum probability, to categorize each image. This strategy achieves an average 98.7% accuracy on the validation dataset (Lecun et al. 1998).

*Retinal transform: Peripheral vision and log Polar encoding* The non-uniform sampling of visual space is adequately modeled as a log-polar conformal mapping (Javier Traver and Bernardino 2010) which has a long history in computer vision and robotics. A first property of this mapping is the separation between the foveal and the peripheral areas as we defined above. This transformation has also other notable properties, such as the correspondence by way of translations in the radial and angular directions to respectively rotations and scalings in the visual domain. However, this sensor is to our knowledge most often not coupled to an action.

First, both the visual features and the expected target position may to be expressed in retinal coordinates which we choose here to be log-polar as it provides a good fit with observations in mammals (Javier Traver and Bernardino 2010). On the visual side, we extracted local visual features as oriented edges as the combination of the retinotopic transform with that of the primary visual cortex (Fischer et al. 2007). The centers of these first and second order orientation filters are radially organized around the center of fixation, with small and tightened receptive fields at the center and more large and scarce receptive fields at the periphery, see Figure 2-B. The size of the filters increases proportionally to eccentricity. The filters are organized in 10 spatial eccentricity scales (respectively placed at around 2, 3, 4.5, 6.5, 9, 13, 18, 26, 36.5 , and 51.3 pixels from the center) and 16 different azimuth angles allowing them to cover most of the original 128 × 128 image. At each of these position, we computed 10 different edge orientations and 2 different phases (symmetric and anti-symmetric) using log-Gabor filters (Fischer et al. 2007). This finally implements a (fixed) bank of linear filters which model the receptive fields of the input to the primary visual cortex.

From any input image ($128 \times 128 = 16384$ pixels) is linearly transformed into a retinal activity vector $\boldsymbol{x}$. Note that to ensure the balance of the coefficients across scales, the images are first whitened. The length of this vector is 1600 such that the retinal filter compresses the original image by about 90%, with high spatial frequencies preserved at the center and only low spatial frequencies conserved at the periphery. In practice, this filters are pre-computed and placed into a matrix for a rapid transform of batches of input displays into retinal transforms. This matrix transformation allows also the evaluation of a reconstructed visual image given a retinal activity vector thanks to the pseudo-inverse matrix of the forward transform matrix. In summary, the full-sized images are transformed into a peripheral retinal image which will be fed to the "where" pathway.

*Collicular representation: accuracy map* The output of the “Where” pathway is defined as an *accuracy map* representing the probability of the presence of a target in the visual field, independently of its identity. As the retinotopic map, this target accuracy map is also organized radially in a log-polar fashion, making the target position estimate more precise at the center and fuzzier at the periphery. This modeling choice is reminiscent of the approximate log-polar organization of the superior colliculus (SC) motor map. In ecological conditions, the accuracy map is trained by sampling, i.e. by "trial and error", using for instance corrective saccades to compute (a posteriori) the probability of a correct localization. In a computer simulation however, this induces a combinatorial explosion which does render the calculation not amenable.

However, as we generated the display, we know the position of the target (which is hidden to the agent). Moreover, we observed that we could also evaluate the accuracy of the classifier (that is, of the fixed “what” pathway) knowing the translational shift imposed to the input foveal image by a saccade of known amplitude. Knowing the size of the $28 \times 28$ input image, this generates a $55 \times 55$ accuracy map (larger shift correspond to a target outside the fovea and thus an accuracy corresponding to the chance level of 10%). This classifier displays a high accuracy at the center (with a value of 98.7% corresponding to the validation score without any translational shift), and a fast decreasing accuracy with target eccentricity. By assuming ergodicity, knowing the centered accuracy map allows to rapidly predict for each visual sample the full accuracy map at each pixel by shifting the centered accuracy map on the true position of the target. Such a computational shortcut is allowed by the independence of the categorical performance with position.

Finally, this full accuracy map is log-polar projected to provide the expected accuracy of each hypothetical saccade in a retinotopic space (see Figure 2-B). In practice, we use the energy of the filters at each position as a proxy to quantify the projection from the metric space of the display to the retinotopic space. This generates a filter bank at 10 spatial eccentricity scales and 16 different azimuth angles, and 160 output filters. Each filter is normalized such that the value at each log-polar position is the average of the values which are integrated in visual space. Applied to the full sized ground truth accuracy map computed in metric space, this gives an accuracy map at different location of a retinotopic motor space. Such transform is again implemented by a simple matrix multiplication which can be precomputed to fasten calculations. Practically, this also allows to compute an inverse transform using the pseudo-inverse matrix of the forward transform. In particular, we use that inverse transform to represent the accuracy predicted by any given log-polar vector, but also to compute the position of maximal accuracy in metric space to set up the sensor displacement.

### 3.3 Classifier training using deep learning

Modern parametric classifiers are composed of many layers (hence the term “Deep Learning”) that can be trained through gradient descent over arbitrary input and output feature spaces. The ease of use of those tightly optimized training algorithms allows for the quantification of the difficulty of a task through the failure or success of such training. Consider the retinal transform $\boldsymbol{x}$ as the input and a log-polar retinotopic vector $\boldsymbol{a}$ made of $n$ Bernouilli probabilities (success probabilities) as the output. The network is trained to predict the distribution $\boldsymbol{a}$ knowing the retinal input $\boldsymbol{x}$ by comparing it to the known ground truth distribution computed over the motor map. As a loss function, we will naturally use the Kullback-Leibler divergence between the ground truth and the predicted map.

In practice, the parametric neural network is made of an input (retinal) layer, two fully connected hidden layers of size 1000 and an output layer, with ReLu

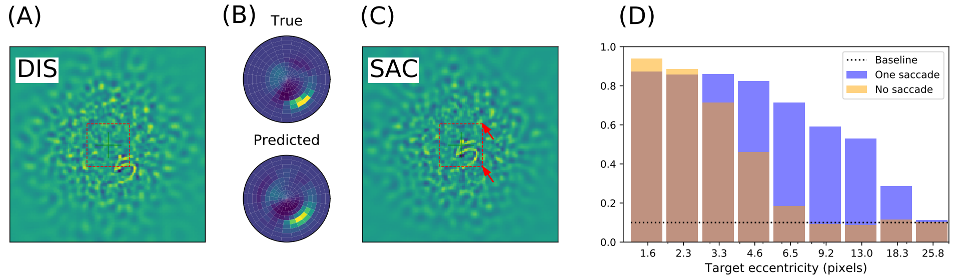

**Fig. 3. Simulated active vision agent**: **(A)** The visual display (`DIS`, see also Figure 1-C) is transformed into a retinotopic representation which is used as the input of a multi-layer neural network implementing the “where” pathway, that transforms the retinal image into an accuracy map. **(B)** We show after training a typical network output ('Predicted') as compared with the ground truth ('True'). **(C)** The network output allows to generate a saccade to the most likely target position in visual space and to recenter the retinotopic map (`SAC`), making possible to estimate a final classification rate. **(D)** The active vision agent is tested at different eccentricities (in pixels). Orange bars: accuracy of a central classifier ('No saccade') with respect to the target's eccentricity, averaged over 1,000 trials per eccentricity scale. Blue bars: Final classification rate after one saccade predicted by the “Where” pathway.

activation between each layer, except at the output which uses a sigmoid function to ensure that the output is compatible with the representation of a likelihood (see Figure 2-C). Another improvement in convergence speed that was obtained by using batch normalization. The network is trained over $500,000$ saccades on full-images, using the binary cross-entropy loss as the error signal, with a learning rate equal to $10^{-4}$ and stochastic gradient descent with a momentum to improve convergence. The training is done for 25 epochs in about 1 hours on a laptop. The code is written in Python (version 3.7.6) with pyTorch library (Paszke et al. 2017) (version 1.0.1). The full scripts for reproducing the figures and extending the results to a full range of parameters is available at `https://github.com/laurentperrinet/WhereIsMyMNIST`.

## 4 Results

### 4.1 Inferring where to look

After training the network, we evaluated simulations of the final accuracy at the landing of the predicted saccade (see Figure 3). For each different visual display (a different digit at a different position with a different noise clutter), a retinocentric visual input is processed (figure 3-A), providing a predicted accuracy map (figure 3-B) that can be compared to the actual future accuracy. Then, a saccade is carried out based on the most probable position as computed from the predicted accuracy map (figure 3-C), and the final accuracy is computed from the “what” pathway using LeNet model. This is repeated $1,000$ times at different eccentricities, and the final average accuracy is shown as blue bars on figure 3-D. It is compared to a central classifier trying to predict the category without doing a saccade (orange bars). As expected, the accuracy decreases with the eccentricity, for the targets become less and less visible in the periphery. The decrease is very rapid in the central classifier case: the accuracy drops to the baseline level after the fourth scale, which corresponds to a 4.6 pixels radius around the center of fixation). In contrast, issuing a saccade is beneficial in up to 18 pixels around the fixation center, allowing a much wider covering of the initial image. The difference between the two distributions forms an "accuracy gain", that quantifies the benefit of active inference with respect to a central prior, interpreted as the information gain provided by the "Where" pathway.

As our saccade selection algorithm may implement the essential operations done in the "Where" pathway, the central classifier may also reflect the response of the "What" pathway, giving the potential category of the digit. It is therefore possible to compare the two accuracy estimates to chose the most appropriate action: it may be that the accuracy is best in the "What" pathway and in that case no saccade is produced. The decision frontier lies between the first and the second spatial scale, allowing to pursue micro-saccades in the close vicinity of the target (2-3 pixels), in order to achieve a perfect centering. In the other decision case, the ”What" accuracy can still be considered to update the "Where" accuracy. This allows in particular to "explain away" the current position of the fixation and the neighboring ones. Such heuristic gives a principled formulation of the inhibition of return mechanism which is an important aspect for modeling saccades (Itti and Koch 2001). In particular, we predict that such a mechanism is dependent on the class of inputs, and would be different for searching for faces as compared to digits. In addition, note that these results are robust to changes in experimental or network parameters (see Figure 4).

### 4.2 Quantitative role of parameters

## 5 Discussion

### 5.1 Summary

The predicted accuracy map has, in our case, the role of a value-based action selection map, as it is the case in model-free reinforcement learning. However, it also owns a probabilistic interpretation that may be combined with other accuracy predictions (such as the one done through the "what" pathway) which potentially explains more elaborate decision making, such as inhibition of return. The approach is also energy-efficient as it includes the strong compression rate performed by retina and V1 encoding, which is preserved up to the action selection level. When energy and computing power are at stake, as it is the case in bio-inspired robotics, it may thus be relevant to envision our implementation. By preserving a probabilistic interpretation in bio-realistic action selection, we also allow for a principled use of human visual scan path over images, such as is generally modeled as low-level feature-based saliency maps (Itti and Koch 2001). It may indeed be possible to consider the actual action selection as implementing focal accuracy-seeking policy across the image, and learn the actual saccade path as the response of a pre-trained accuracy prediction. Identified regions of interest may then be compared with the baseline bottom-up approaches.

### 5.2 Limits

From the modeling side, our model still relies on a strong idealization, assuming the existence of a probabilistic representation of action achievement over large action spaces in the brain. Similarly, how the brain may combine and integrate various probabilities is still an open question, that resorts to the classical binding problem. At last, the presence of many targets in a scene should also be addressed by the model, which resorts to sequentially select targets, in combination with a concrete implementation of an inhibition of return mechanism.

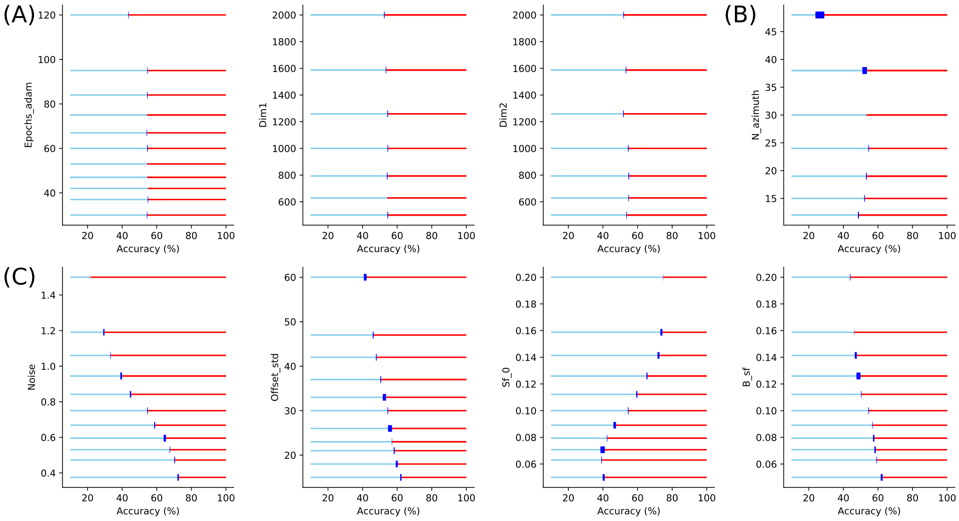

**Fig. 4. Quantitative role of parameters**: We show here variations of the average accuracy as a function of some free parameters of the model. All parameters of the presented model were tested, from the architecture of image generation, to the parameters of the neural network implementing the "Where" pathway (including meta-parameters of the learning paradigm). We show here the results which show the most significative impact on average accuracy. **(A)** First, we scanned parameters of the Deep Learning neural network. It shows that accuracy quickly converged after a characteristic time of approximately 25 `epochs`. We then tested different values for the dimension of respectively the first (`dim1`) and second (`dim2`) hidden layers, showing marginal changes in accuracy. **(B)** The accuracy also changes with the architecture of the foveated input as shown here by changing the number `N_azimuth` of azimuth directions which are sampled in visual space. This shows a compromise between a rough azimuth representation and a large precision, which necessitates a longer training phase, such that the optimal number is around 20 azimuth directions. **(C)** Finally, we tested some properties of the input, respectively from left to right: noise level (`noise`), mean spatial frequency of clutter `sf_0` and bandwidth `B_sf` of the clutter noise. This shows that average accuracy evolves with noise (see also Figure 3 for an evolution as a function of eccentricity), but also to the characteristics of the noise clutter. In particular, there is a drop in accuracy whenever noise is of similar wavelength as digits, but which becomes less pronounced as the bandwidth increases.

## References

Butko, Nicholas J and Javier R Movellan (2010). "Infomax control of eye movements". In: *IEEE Transactions on Autonomous Mental Development* 2.2, pp. 91–107.

Fischer, Sylvain et al. (Jan. 2007). "Self-Invertible 2D Log-Gabor Wavelets". In: *International Journal of Computer Vision* 75.2, pp. 231–246. DOI: `10.1007/s11263-006-0026-8`.

Friston, Karl J et al. (2012). "Perceptions as Hypotheses: Saccades as Experiments". In: *Frontiers in Psychology* 3. DOI: `10.3389/fpsyg.2012.00151`.

He, Kaiming et al. (Feb. 2015). "Delving Deep into Rectifiers: Surpassing Human-Level Performance on ImageNet Classification". In: *arXiv:1502.01852 [cs]*. 04208.

Itti, Laurent and Christof Koch (2001). "Computational Modelling of Visual Attention". In: *Nature Reviews Neuroscience* 2.3, pp. 194–203. DOI: `10/chw2bk`.

Javier Traver, V and Alexandre Bernardino (2010). "A review of log-polar imaging for visual perception in robotics". In: *Robotics and Autonomous Systems* 58.4, pp. 378–398.

Kirchner, H and Sj Thorpe (2006). "Ultra-Rapid Object Detection with Saccadic Eye Movements: Visual Processing Speed Revisited". In: *Vision Research* 46.11, pp. 1762–76. DOI: 10.1016/j.visres.2005.10.002.

Kümmerer, Matthias, Thomas S. A. Wallis, and Matthias Bethge (Oct. 2016). "DeepGaze II: Reading Fixations from Deep Features Trained on Object Recognition". In: *arXiv:1610.01563 [cs, q-bio, stat]*. 00075. arXiv: 1610.01563 [cs, q-bio, stat].

Lecun, Y. et al. (Nov. 1998). "Gradient-Based Learning Applied to Document Recognition". In: *Proceedings of the IEEE* 86.11. 14279, pp. 2278–2324. DOI: 10/d89c25.

Mirza, M. Berk et al. (Jan. 2018). "Human visual exploration reduces uncertainty about the sensed world". In: *PLOS ONE* 13.1. Ed. by Stefan Kiebel, e0190429. DOI: 10.1371/journal.pone.0190429.

Najemnik, Jiri and Wilson S. Geisler (2005). "Optimal Eye Movement Strategies in Visual Search". In: *Nature* 434.7031, pp. 387–391. DOI: 10/bcbw2b.

Paszke, Adam et al. (Oct. 28, 2017). "Automatic Differentiation in PyTorch". In: 00918.

Redmon, Joseph et al. (June 2015). "You Only Look Once: Unified, Real-Time Object Detection". In: 03448.

Ren, S. et al. (June 2017). "Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks". In: *IEEE Transactions on Pattern Analysis and Machine Intelligence* 39.6. 08051, pp. 1137–1149. DOI: 10/gc7rmb.

Samonds, Jason M., Wilson S. Geisler, and Nicholas J. Priebe (Nov. 2018). "Natural Image and Receptive Field Statistics Predict Saccade Sizes". En. In: *Nature Neuroscience* 21.11. 00002, p. 1591. DOI: 10/gfgt3k.

Sandler, Mark et al. (Jan. 2018). "MobileNetV2: Inverted Residuals and Linear Bottlenecks". In: *arXiv:1801.04381 [cs]*. 00241.

Sanz-Leon, Paula et al. (Mar. 2012). "Motion Clouds: Model-Based Stimulus Synthesis of Natural-like Random Textures for the Study of Motion Perception". In: *Journal of Neurophysiology* 107.11, pp. 3217–3226. DOI: 10.1152/jn.00737.2011.

Treisman, Anne M and Garry Gelade (1980). "A Feature-Integration Theory of Attention". In: *Cognitive psychology* 12.1. 11957, pp. 97–136. DOI: 10.1016/0010-0285(80)90005.

Võ, Melissa L.-H. and Jeremy M. Wolfe (2012). "When Does Repeated Search in Scenes Involve Memory? Looking at versus Looking for Objects in Scenes". In: *Journal of Experimental Psychology: Human Perception and Performance* 38.1, pp. 23–41. DOI: 10.1037/a0024147.