A dual foveal-peripheral visual processing model implements efficient saccade selection

Emmanuel Daucé[1], Pierre Albiges[2], Laurent Perrinet[2,*]

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

* Laurent.Perrinet@univ-amu.fr

Abstract

We develop a visuo-motor model that implements visual search as a focal accuracy-seeking policy across a crowded visual display. Stemming from the active inference framework, saccade-based visual exploration is idealized as an inference process, assuming that the target position and category are independently drawn from a common generative process. This independence allows to divide the visual processing in two independent pathways, consistently with the anatomical "What"/"Where" separation. A biomimetic log-polar treatment of the visual information, that includes the strong compression rate performed at the sensor level by retina and V1 encoding, is preserved up to the action selection level. A dual neural network architecture, that independently learns where to look and what to see, is then trained, with the foveal accuracy used as a monitoring signal for action selection. This allows in particular to interpret the "Where" as a retinotopic action selection pathway, that drives the fovea toward the target position, in order to increase the recognition accuracy. A specific approximate Information Gain metric, taken as the difference between central and peripheral accuracy, is used for action selection after training. The comparison of both accuracies amounts either to select a saccade or to keep the eye focused at the center, so as to identify the target. Tested on a simple task of finding digits in a large, cluttered image, simulation results demonstrate the benefit of our approach, whose key computational shortcuts finally provide ways to implement visual search in a sub-linear fashion, in contrast with mainstream computer vision.

Author summary

The visual search task consists in extracting a scarce and specific visual information (the "target") from a large and crowded visual display. In computer vision, 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. Then, visual processing is separated in two specialized pathways, the "where" pathway mainly conveying information about target position in peripheral space (independently of its category), and the "what" pathway mainly conveying information about the category of the target (independently of its position). This object recognition pathway is shown here to have an essential role, providing an "accuracy drive" that serves to force the eye to foveate peripheral objects in order to increase the peripheral accuracy, much like in the "actor/critic" framework.

Put together, all those principles to provide ways toward both adaptive and resource-efficient visual processing systems.

Introduction

Problem statement.

Past 10 years have seen the disrupting development of deep learning based image processing. Indeed the field of computer vision has been recast by the outstanding capability of convolution-based deep networks to capture the semantic content of images and photographs. Image processing algorithms recently outreached the performance of human observers in specific image categorization tasks [**?**]. Their success relies on a reduction of parameter complexity through weight sharing in convolutional neural networks applied over the full image. Initially trained on energy greedy, high performance computers, they are now designed to work on more common hardware such as desktop computers with dedicated GPU hardware [**?**]. However, despite lot of efforts spent in optimizing the processing costs, the processing of large images is still done at a cost that scales linearly with the image size. All regions, even the "boring" ones are systematically scanned and processed in parallel through dedicated hardware at a significant computational cost. Image processing architectures consequently contain millions of parameters with subsequent energy consumption while still handling relatively small images. This introduces a trade-off between efficiency and accuracy, for instance in autonomous driving, with the need to detect visual objects at a glance while running on resource-constrained embedded hardware.

In contrast, when human vision is considered, things work differently. First, the general performance is still greater than that of computer vision. Indeed, object recognition can be achieved by the human visual system both rapidly, – in less than 100 ms [**?**] – 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. 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 human vision is *dynamic*. 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. The scanning of a full visual scene is thus not done in parallel but sequentially, and only scene-relevant regions of interest are scanned through saccades. This implies a *decision process* at each step that decides *where to look next*. 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.

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 face despite the sensory clutter in the visual field. 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 identify a particular object *class* (e.g. human faces) given all their possible spatial configurations and respective geometrical visual transformations. Searching for *any* face in a peripheral and crowded display

needs to precede the recognition of a specific face identity.

State of the art

To take benefit from this visuomotor behavior, it is important 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. Addressing apparently simple questions such as "find the green bottle on the table", it 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. When restricted to a mere "feature search" [**?**], many solutions are proposed. Notably, recent advances in deep-learning have provided efficient models such as faster-RCNN [**?**] or YOLO [**?**]. Their object search implementations predict in the image the probability of proposed bounding boxes around visual objects. While rapid, the number of boxes may significantly increase with image size and the approach more generally necessitates dedicated hardware to run in real time.

In parallel, human visual scan-path over natural images provide ways to define *saliency maps*, that quantify the attractiveness of the different parts of an image, that are consistent with the detection of objects of interest. Essential to understand and predict saccades, they also serve as phenomenological models of attention. Estimating the saliency map from a luminous image is a classical problem in neuroscience, that was shown consistent with a distance from baseline image statistics known as the "Bayesian surprise" [**?**]. The saliency approach was recently updated using deep learning to estimate saliency maps over large databases of natural images [**?**]. 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 the active vision process.

Foveated models of vision have been considered for long time in robotics and computer vision as a way to leverage the visual scene scaling problem. Focal image processing relies a non-homogeneous compression of an image, that maintains the pixel information at the center of fixation and strongly compresses it at the periphery, including pyramidal encoding [**?**, **?**], local wavelet decomposition [**?**] and logpolar encoding [**?**, **?**]. Though focal and multiscale encoding is now largely considered in static computer vision, sequential implementations have not been shown effective enough to overtake static object search methods. Several implementations of a focal sequential search in visual processing can be found in the literature, with various degrees of biological realism [**?**, **?**], that often rely on a simplified focal encoding, long training procedures and bounded sequential processing. More realistic attempts to combine foveal encoding and sequential visual search can be found in [**?**, **?**, **?**], that will be compared further on with our approach.

In contrast to phenomenological (or "bottom-up") approaches, active models of vision [**?**, **?**, **?**] provide the ground principles of saccadic exploration. In general, 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 [**?**]. 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 speed. Looking more closely at neurophysiology, the study of [**?**] 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.

A further modeling perspective is provided by [**?**]. In this setup, a full description of the visual world is used as a generative process. An agent is completely described by 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 [**?**] 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 principles, our aim is to produce a principled and resource-effective model of vision. We start from an elementary visual search problem, that is how to locate an object in a large, crowded image, and take human vision as a guide for efficient design. Our framework is made as general as possible, with minimal mathematical treatment, to speak largely to fragmented domains, such as machine learning, neuroscience and robotics. We expect to provide an integrated view of foveated active vision, applicable to both domains.

After this introduction, the principles underlying accuracy-based saccadic control are defined in the second section. We first define notations, variables and equations for the generative process governing the experiment and the generative model for the active vision agent. Complex combinatorial inferences are here replaced by separate pathways, i.e. the spatial ("Where") and categorical ("What") pathways, whose output is combined to infer optimal eye displacements and subsequent identification of the target. Our agent, equipped with a foveated sensor, should learn an optimal behavior strategy to actively scan the visual image. Implementation details are provided in the methods section, giving ways to reproduce our results, showing in particular how to simplify the learning using accuracy-driven action maps. Numerical simulations are presented in the results section, demonstrating the applicability of this framework to different task complexity levels. The last section finally summarizes the results, showing its relative advantages in comparison with other frameworks, and providing ways toward possible improvements.

Principles

For biological vision is the result of a continual optimization under strong material and energy constraints, we need to understand both its ground principles and its specific computational and material constraints in order to implements effective biomimetic vision systems.

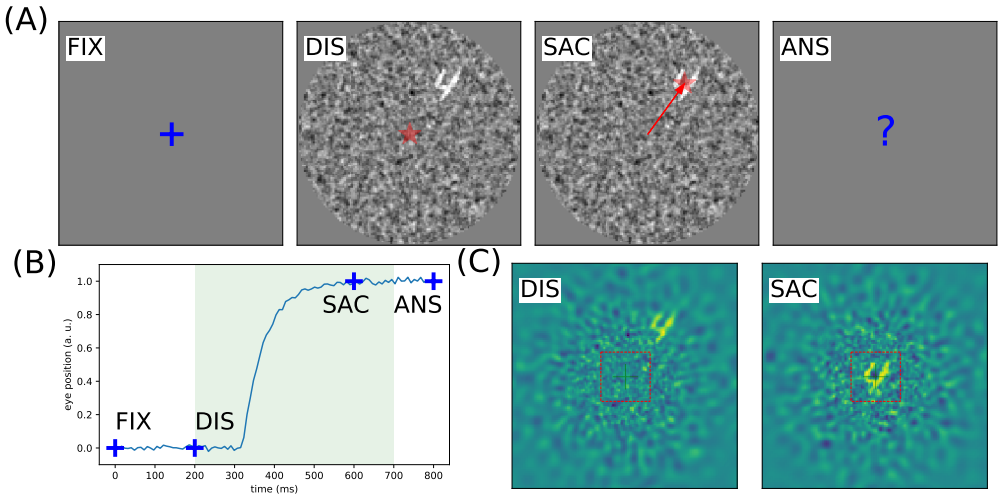

Fig 1. Problem setting: In generic, ecological settings, the visual system faces a tricky problem when searching for one target (from a class of targets) in a cluttered environment. It is synthesized in the following experiment: **(A)** After a fixation period `FIX` of 200 ms, an observer is presented 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 on the potential target (`SAC`, here successful). Finally, the observer has to identify the digit by a keypress `ANS`. *NB*: the target contrast is here enhanced for a better readability. **(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 foveal region. 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 contrast 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.

In order to do so, we provide a simplified visual environment toward which a visual agent can act on. The search experience is formalized and simplified in a way reminiscent to classical psychophysic experiments: an observer is asked to classify digits (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). In order to vary the difficulty of the task, different parameters are controlled, such as the target eccentricity, the background noise period and and the signal/noise ratio (SNR). The agent initially fixates the center of the screen. Due to the peripheral clutter, he needs to explore the visual scene through saccades to provide the answer. He controls a foveal visual sensor that can move over the visual scene through saccades (see Figure 1-B). When a saccade is actuated, the center of fixation moves toward a new location, which updates the visual input (see Figure 1-C). The lower the SNR and the larger the initial target eccentricity, the more difficult the identification. There is a range of eccentricities for which it is impossible to identify the target from a single glance, so that a saccade is

necessary to issue a proper response. This implies in general that the position of the object may be detected in the first place in the peripheral clutter before being properly identified.

This setup provides the conditions for a separate processing of the visual information. Indeed, in order to analyze a complex visual scene, there are two types of processing that need to be done. On the one side, you need to analyze in detail what is at the center of fixation, that is the region of interest currently processed. On the other side, you also need to analyze the surrounding part, even if the resolution is low, in order to choose what is the next center of fixation. This basically means making a choice of "what's interesting next". You do not necessarily need to know what it is, but you need to that it's interesting enough, and of course you need to know what action to take to move the center of fixation at the right position. This is reminiscent of the What/Where separate visual processing separation observed in monkeys and humans ventral and dorsal visual pathways [**?**].

Active inference

This kind of reasoning can be captured by a statistical framework called a partially observed Markov Decision Process (POMDP), where the cause of a visual scene is couple made of a viewpoint and scene elements. Changing the viewpoint will conduct to a different scene rendering. A generative model tells how typically looks the visual field knowing the scene elements and a certain viewpoint. In general, 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. Here 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, or "viewpoint") and y the object shape (or object identity). The visual field x is the state of the sensors, that is, a partial view of the visual scene, measured through the generative process : $x \sim p(X|e)$.

Using Bayes rule, one may then infer the scene elements from the current view point (model inversion). The real physical state e being hidden, a parametric model θ is assumed to allow for an estimate of the cause of the current visual field through model inversion thanks to Bayes formula, in short:

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

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. Actuating a command a changes the viewpoint to u', which feeds the system with a new visual sample $x' \sim p(X|u', y)$. The more viewpoints you have, the more certain you are about the object identity through a chain rule sequential evidence accumulation.

In an optimal search setup however [**?**], you need to choose the next viewpoint that will help you *the most* to disambiguate the scene. 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 simplified setups, and thus infeasible in practice.

The predictive approach 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 independence of the factors that are the cause of the sensory view. The independence hypothesis allows considering the viewpoint 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 [**?**] for more complex visual scene grammars).

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. 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. 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 yet be too 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. To sum up, a main assumption here is that instead of trying to detect the actual position of the target, better off for the agent is to estimate how accurate the categorical classifier will be after moving the eye. Extended to the full action space A, 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. Our main assumption here is that such a *predictive accuracy map* is at the core of a realistic saccade-based vision systems. Compared with a baseline approach that would predict for all possible gaze directions over an image, this map should moreover be organized radially to preserve the retinotopic compression.

Finally, the independence assumption allows to separate the scene analysis in two independent tasks. Each task is assumed to be realized in parallel through distinct computational pathways, that will be referred as the "What" and the "Where" pathways by analogy with the ventral and dorsal pathways in the brain (see figure 2). 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 density of cones, and thus higher spatial precision. In contrast, the saccade planning should rely on the full visual field, with peripheral regions having a lower sensor density and a lesser sensitivity to high spatial frequencies. 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 2C). Accordingly with the visual 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

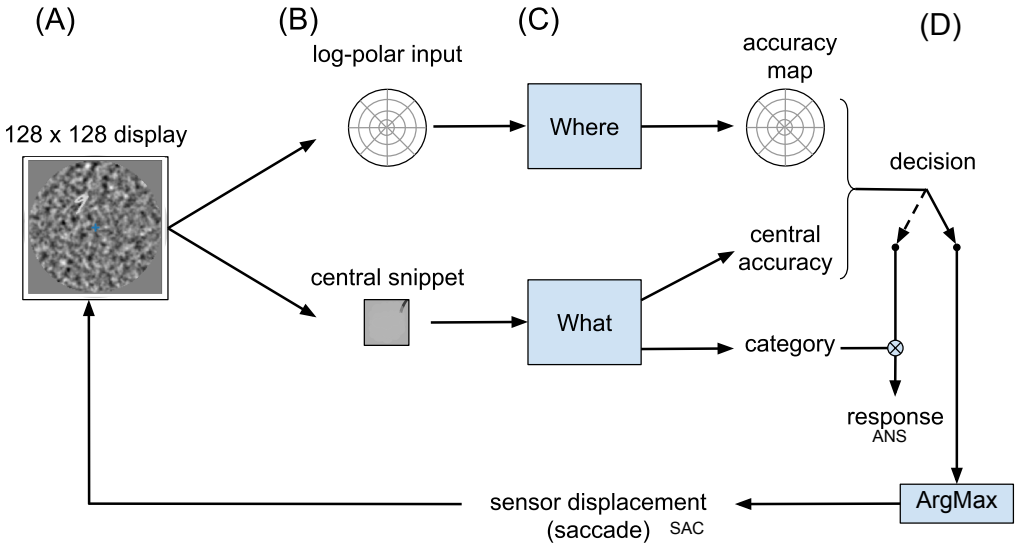

Fig 2. Computational graph. Two streams of information are separated from the visual primary layers, one stream for processing the central pixels only, the other for processing the periphery with a logpolar encoding. The two streams converge toward a decision layer that compares the central and the peripheral acuracy, in order to decide wether to issue a saccadic or a categorical response. If a saccade is produced, then the center of vision is displaced toward the region that shows the higher accuracy on the accuracy map. **(A)** The visual input is constructed the following way: first a 128×128 natural-like background noise is generated, characterized by noise contrast, mean spatial frequency and bandwidth [**?**]. Then a circular mask is put on. Last a sample digit is selected from the MNIST database (of size 28×28), rectified, multiplied by a contrast factor and overlayed on the background at a random position (see an example in Figure 1-A, `DIS`). **(B)** The visual input is then transformed in 2 ways: (i) a 28×28 central foveal-like snippet is fed to a classification network ("What" pathway) and (ii) a log-polar set of oriented visual features is fed to the "Where" pathway. This log-polar input is generated by a bank of filters whose centers are positioned on a log-polar grid and whose radius increases proportionally with the eccentricity. **(C)** The "What" network is implemented using the three-layered LeNet CNN [**?**], while the "Where" network is implemented by a three-layered neural network consisting of the retinal input, two hidden layers with 1000 units each and a collicular-like accuracy map at the output. This map has a similar retinotopic organization and predicts the accuracy of each hypothetical position of a saccade. 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. **(D)** If the predicted accuracy in the output of the "Where" network is higher than that predicted in the "What" network, the position of maximal activity in the "Where" pathway serves to generate a saccade which shifts the center of gaze.

region responsible for oculo-motor control [**?**]. The saccadic motor output showing a similar log-polar compression than the visual input, the saccades should be more precise at short than at long distance (and several saccades may be necessary to precisely reach distant targets).

Detailed implementation

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 the training. The simplified anatomy of the agent is composed of two separate pathways whose processing is realized by such a neural network. Each network is trained and tested separately on distinct datasets, before being finally evaluated in a dynamic vision setup (see next section).

Images generation

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 [**?**]: Samples are drawn from the database of 60000 grayscale 28×28 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 128×128. 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 [**?**] 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. The spatial frequency is optimized at 0.1 pixel^{-1} to fit that of the original digits. This specific spatial frequency occasionally allows to generate some "phantom" digit shapes in the background. 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.

Foveal vision and the "What" pathway

At the core of the vision system is the identification module, i.e. the "What" pathway. It consists of a classic convolutional classifier showing some translation invariance. This translation invariance can be measured in the form of a shift-dependent accuracy map. Importantly, it can quantify its own classification uncertainty, that may allow comparisons with the output of the "Where" pathway.

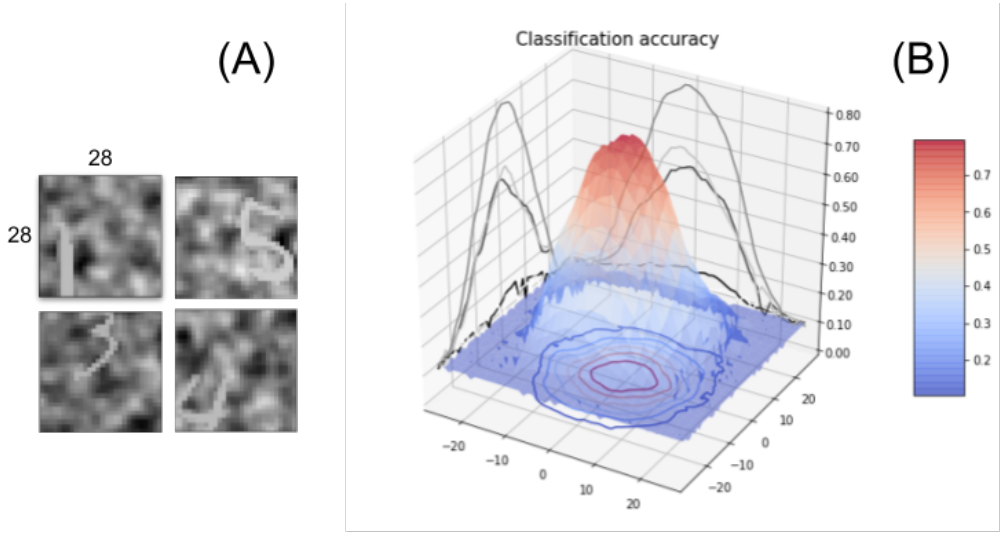

Fig 3. **(A)** Input samples from the "What" training set, with randomly shifted targets using a Gaussian bivariate spatial offset with a standard deviation of 15 pixels. The target contrast is randomly set between 0.3 and 0.7. **(B)** 55×55 shift-dependent accuracy map, measured for different target eccentricities on the test set after training.

The foveal input is defined as the 28×28 grayscale image extracted at the center of gaze (see dashed red box in Figure 1-C). This image is passed unmodified to the agent's visual categorical pathway (the "What" pathway), that is realized by a convolutional neural network, here the known "LeNet" classifier [**?**]. The network structure, that processes the input to identify the target category, is provided (and unmodified) by the pyTorch library [**?**]. It is made of a 3 convolution layers followed by two fully-connected layers. The network output is a vector representing the probability of detecting each of the 10 digits. The argument of the output neuron with maximum probability provides the image category.

A specific dataset is constructed to train the network. It is made of randomly shifted/randomly attenuated digits overlayed over a noisy background, as defined above. Both the offset, the contrast and the background noise render the task more difficult than the original MNIST classification. The relative contrast of the digit is randomly set between 0.3 and 0.7. The network is trained incrementally by progressively increasing the offset variability (of a bivariate central gaussian) by increasing the standard deviation from 0 to 15 (with a maximal offset set at 25 pixels). The network is trained on a total of 75 epochs, with 60000 examples generated at each epoch from the MNIST original training set. The shifts and backgrounds are re-generated at each epoch. The shift standard deviation increases of one unit every 5 epochs. Note that at the end of the training, many digits fall outside the center of the fovea, so that many examples are close to impossible to classify, either because of a low contrast or a too large eccentricity. At the end of the training process, the average accuracy is thus of 34% (though it had a 91% accuracy after the 5th epoch, when the digits were only at the center).

After training, a shift-dependent accuracy map is computed by systematically testing the network accuracy on every horizontal and vertical offset, each on a set of 1000 samples generated from the MNIST test set, within a range of $+/-27$ pixels (see figure 3). This forms a 55×55 accuracy map showing higher accuracy at the center, and a slow decreasing accuracy with target eccentricity (with over 70% accuracy plateau showing a shift invariance on a 7-8 pixels eccentricity radius). This significant shift invariance is a known effect of convolutional computation, that is obtained here at the

cost of a lesser central recognition rate (around 80%), remembering the classification task is here harder by construction. The accuracy fastly drops for greater than 10 pixels eccentricity, reaching the baseline 10% chance level at around 20 pixels offset.

Peripheral vision: from log-polar feature vectors to log-polar action maps

The "Where" pathway is devoted to choosing the next saccade. Here we assume the "Where" implements the following action selection: where to look next in order to reduce the uncertainty about the target identity? This implies moving the eye such as to increase the "What" classifier accuracy. For a given visual field, each possible future saccade has an expected accuracy, that can be trained from the "What" pathway output. To accelerate the training, we use a shortcut that is training the network on a translated accuracy map. The ouput is thus an accuracy map, that tells for each possible visuo-motor displacement the value of the future accuracy.

Primary visual representation: log-polar orientation filters For to reduce the processing cost, and in accordance with observations [**?**, **?**], a similar log-polar compression pattern is assumed to be conserved from the retina up to the primary motor layers. The non-uniform sampling of the visual space is adequately modeled as a log-polar conformal mapping, as it provides a good fit with observations in mammals [**?**] which has a long history in computer vision and robotics. Both the visual features and the output accuracy map are to be expressed in retinal coordinates. On the visual side, local visual features are extracted as oriented edges as a combination of the retinotopic transform with primary visual cortex filters [**?**]. 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. The size of the filters increases proportionally to the 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 24 different azimuth angles allowing them to cover most of the original 128×128 image. At each of these position, 6 different edge orientations and 2 different phases (symmetric and anti-symmetric) are computed. This finally implements a (fixed) bank of linear filters which model the receptive fields of the input to the primary visual cortex.

To ensure the balance of the coefficients across scales, the images are first whitened and then linearly transformed into a "primary visual" feature vector $\boldsymbol{x}$. The length of this vector is 2880, such that the retinal filter 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. This matrix transformation allows also the evaluation of a reconstructed visual image given a retinal activity vector thanks to a pseudo-inverse of the forward transform matrix. In summary, the full-sized images are transformed into a primary visual feature vector which is fed to the "Where" pathway.

Visuo-motor representation: "Collicular" accuracy maps The output of the "Where" pathway is defined as an *accuracy map* representing the recognition probability after moving the eye, independently of its identity. Like the primary visual 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, this accuracy map should be trained by

sampling, i.e. by "trial and error", using the actual recognition accuracy (after the saccade) to grade the action selection. In practice, as we generate the visual display, the position of the target (which is hidden to the agent) is known. Under an ergodic assumption, knowing both the translational shift imposed to the visual field by a saccade of known amplitude, and the shift-dependent accuracy map of the "What" classifier (Figure 3-B), the full accuracy map at each pixel can be predicted for each visual sample by shifting the central accuracy map on the true position of the target. Such a computational shortcut is allowed by the independence of the categorical performance with position. This full accuracy map is log-polar projected to provide the expected accuracy of each hypothetical saccade in a retinotopic space. In practice, we use the energy of the filters at each position as a proxy to quantify the projection from the metric space to the retinotopic space. This generates a filter bank at 10 spatial eccentricity scales and 24 different azimuth angles, i.e. 240 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 pre-computed to fasten calculations. Practically, this also allows to compute an inverse transform using the pseudo-inverse matrix of the forward transform. In particular, that inverse transform is used to represent the accuracy predicted by any given visual feature vector, but also to compute the position of maximal accuracy in metric space to set up the sensor displacement.

Classifier 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. The loss function that comes naturally is the Binary Cross-Entropy (negative term of the Kullback-Leibler divergence) between the ground truth and the predicted map (assuming the independence of the output map features).

The parametric neural network consists of a primary visual input layer, followed by two fully connected hidden layers of size 1000 with rectified linear activation, and a final output layer with a sigmoid nonlinearity to ensure that the output is compatible with a likelihood. The network is trained on 60 epochs of 60000 samples, with a learning rate equal to 10^{-4} and the Adam optimizer [**?**]. The full training takes about 1 hours on a laptop. The code is written in Python (version 3.7.6) with pyTorch library [**?**] (version 1.1.0). 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`.

Results

Open loop setup

After training, the "Where" pathway is now capable to predict an accuracy map, whose maximal argument drives the eye toward a new viewpoint. There, a central snippet is extracted, that is processed through the "What" pathway, allowing to predict the digit's label. Examples of this simple open loop sequence are presented in figure 4, when the digits contrast parameter is set to 0.7 and the digits eccentricity varies between 0 and 40 pixels. The presented examples correspond to strong eccentricity cases, when the target is hardly visible on the display (fig. 4a), and almost invisible on the reconstructed input (fig. 4b). The radial maps (fig. 4c-d) respectively represent the

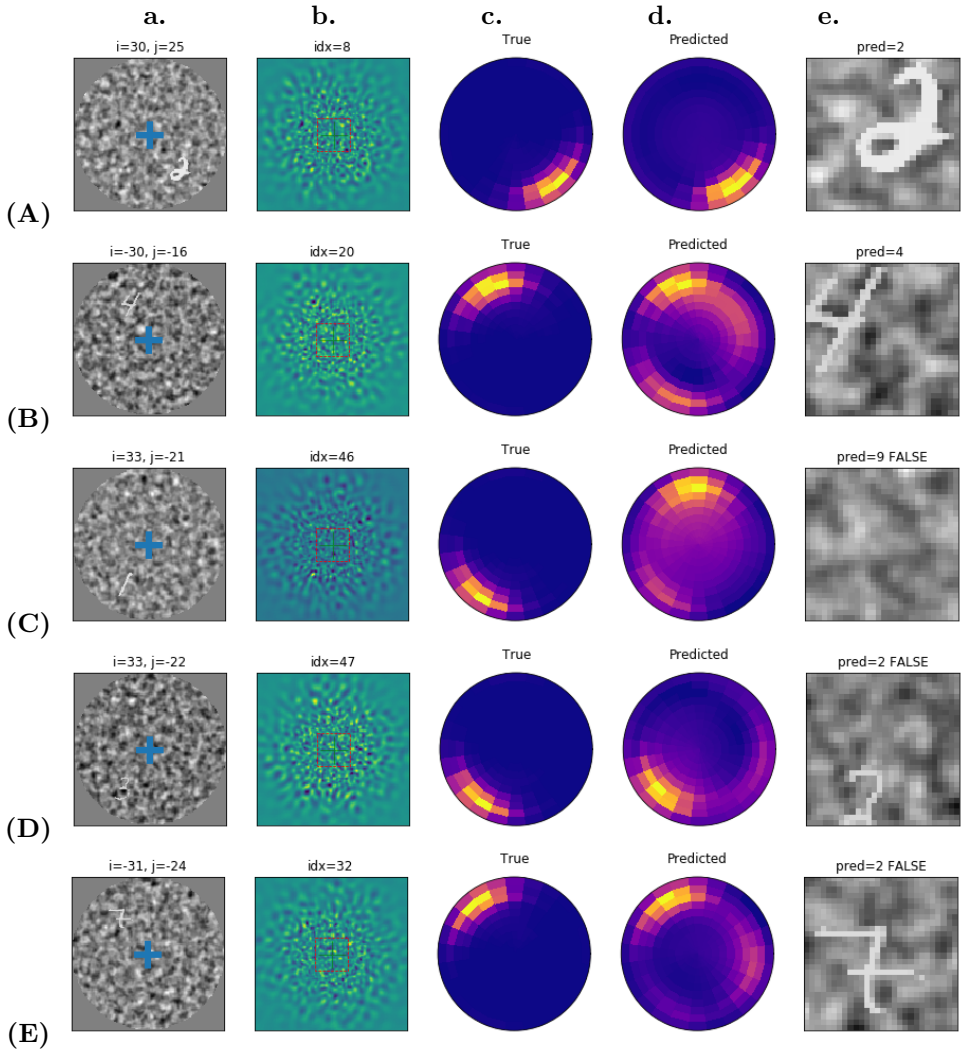

Fig 4. **(A)** – **(E)** Active vision samples after training. **(A)** – **(B)** classification success samples. **(C)** – **(E)** classification failure samples. Digit contrast set to 0.7. From left to right : **a.** The initial 128×128 visual display, with blue cross giving the center of gaze. The visual input is retinotopically transformed and sent to the multi-layer neural network implementing the "Where" pathway. **b.** Magnified reconstruction of the visual input, as it shows off from the primary visual features through an inverse log-polar transform. **c.-d.** Color-coded radial representation of the output accuracy maps, with dark violet for the lower accuracies, and yellow for the higher accuracies. The network output ('Predicted') is visually compared with the ground truth ('True'). **e.** 28×28 central snippet as extracted from the visual display after doing a saccade, with label prediction and success flag in the title.

actual and the predicted accuracy maps. The final focus is represented in fig. 4e, with cases of classification success (fig. 4A-B) and cases of classification failures (fig. 4C-E). In the case of successful detection (fig. 4A-B), the accuracy prediction is not perfect and the digit is not perfectly centered on the fovea. This "close match" however allows for a correct classification for the digit's pixels are fully present on the fovea. The case of fig. 4B and 4C is interesting for it shows two cases of a bimodal prediction, indicating that the network is capable of doing multiple detections at a single glance. The case of

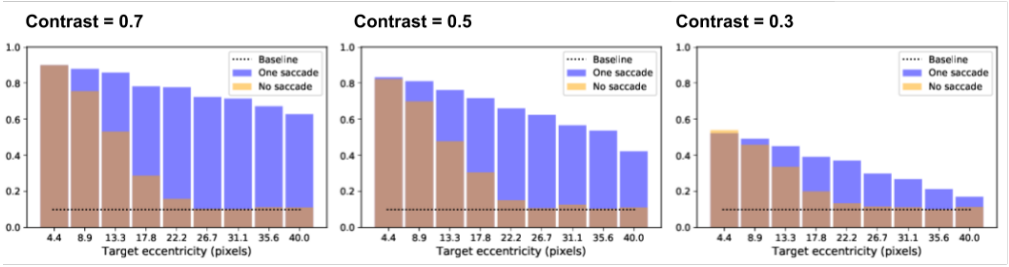

Fig 5. Effect of contrast and target eccentricity. The active vision agent is tested for different target eccentricities (in pixels) and different contrasts to estimate a final classification rate. Orange bars: accuracy of a central classifier ('No saccade') with respect to the target's eccentricity, averaged over 1,000 trials per eccentricity. Blue bars: Final classification rate after one saccade.

4C corresponds to a false detection, with the true target detected still, though with a lower intensity. The case of fig. 4D is a "close match" detection that is not precise enough to correctly center the visual target. Not every pixel of the digit being visible on the fovea, the label prediction is mistaken. The last failure case (fig. 4E) corresponds to a correct detection that is harmed by a wrong label prediction, only due to the "What" classifier inherent error rate.

To test the robustness of our framework, the same experiment was repeated at different signal-to-noise ratios (SNR) of the input images. Both pathways being interdependent, it is crucial to disentangle the relative effect of of both sources of errors in the final accuracy. By manipulating the SNR and the target eccentricity, one can precisely monitor the network detection and recognition capabilities, with a detection task ranging from 'easy" (small shift, strong contrast) to "almost impossible" (large shift, low contrast). The digit recognition capability is systematically evaluated in Figure 5 for different eccentricities and different contrasts. For 3 target contrasts conditions ranging from 0.3 to 0.7, and 10 different eccentricities ranging from 4 to 40 pixels, the final accuracy is tested on $1,000$ trials both on the initial central snippet and the final central snippet (read at the landing of the saccade). The orange bars provide the initial classification rate (without saccade) and the blue bars provide the final classification rate (after saccade) – see figure 5. As expected, the accuracy decreases with the eccentricity, for the targets become less and less visible in the periphery. The decrease is rapid in the central classifier case: the accuracy drops to the baseline level at approximately 20 pixels away from the center of gaze. The saccade-driven accuracy has a much wider range, with a slow decrease up to the border of the visual display (40 pixels away from the center). When varying the target contrast, the initial accuracy profile is scaled by the reference accuracy (obtained with a central target), whose values are approximately 53%, 82% and 92% for SNRs of 0.3, 0.5 and 0.7. The saccade-driven accuracy profile is also similar at the different SNRs values, yet with the scaling imposed by the "What" pathway. This contrast-dependent scaling shows the robustness of our framework to the different factors of difficulty.

The high contrast case (fig. 5A) provides the greatest difference between the two profiles, with an accuracy approaching 0.9 at the center and 0.6 at the periphery. This allows to recognize digits after one saccade in a majority of cases, up to the border of the image, from a very scarce peripheral information. This full covering of the 128×128 image range is done at a much lesser cost than would be done by a systematic image scan, as in classic computer vision. With decreasing target contrast, a general decrease of the accuracy is observed, both at the center and at the periphery, with about 10% decrease with a contrast of 0.5, and 40% decrease with a contrast of 0.3. In addition,

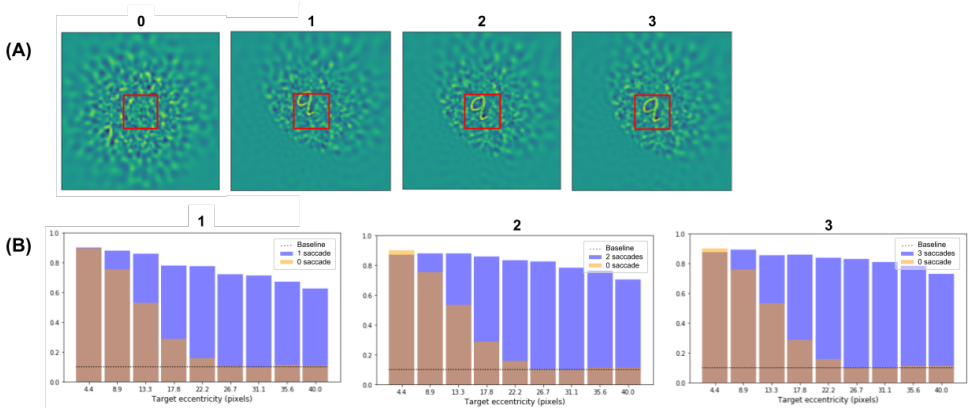

Fig 6. Multi-saccades case. (A) Example of a corrective saccade on a 3-saccades trial. The subjective visual field is reconstructed from the log-polar visual features, with red square delineated 28×28 foveal snippet, after 0, 1, 2 and 3 saccades (from left to right). **(B)** Average classification accuracies measured for different target eccentricities (in pixels) and a different number of saccades. Target contrast set to 0.7. Orange bars: initial central accuracy ('0 saccade') in function of the eccentricity, averaged over 1,000 trials per eccentricity. Blue bars: Final classification rate after one, two and three saccades (from left to right).

the proportion of false detections also increases with contrast decrease. At 40 pixels away from the center, the false detection rate is approximately 30% for a contrast of 0.7, 50% for a contrast of 0.5 and 70% for a contrast of 0.3 (with a recognition close to the baseline at the periphery in that case). The accuracy gain (difference between the initial and the final accuracy) is maximal for eccentricities ranging from 15 to 30 pixels. This optimal range reflects a peripheral region around the fovea where the target detection is possible, but not its identification. The visual agent knows *where* the target is, without exactly knowing *what* it is. More generally, this accuracy difference, that quantifies the benefit of active inference with respect to a central prior, can be interpreted as an approximation of the information gain provided by the "Where" pathway[1].

Closed-loop setup

The most peripheral targets are difficult to detect in one round, resulting in degraded performances at the periphery. Even when correctly detected, our log polar action maps also precludes precise centering. The peripheral targets are generally poorly centered after one saccade, as shown in figure 4, resulting in classification errors. Sequential search is thus needed to allow for a better recognition. Multi-saccades visual search results are thus presented in figure 6

An example of a corrective saccade is shown on figure 6A. A hardly visible peripheral digit target is first approximetely shifted to the foveal zone. A second saccade allows to improve the target centering. A third saccade only marginally improves the centering. As shown in figure 6B, such corrective saccades, that generally only slightly shift the target, still provide a significant improvement in the classification accuracy. Except at the center, the accuracy rises of about 10% both for the mid-range and the most peripheral eccentricities. Most of the improvement however is provided by the first corrective saccade. The second corrective saccade only shows a barely significant 2-3 % improvement, only visible at the periphery. The following saccades would mostly implement target tracking, without providing additional accuracy gain. A

[1] with the true label log-posterior seen as a sample of the posterior entropy – see eq.(1).

3-saccades setup finally allows a wide covering of the visual field, providing a close to central recognition rate at all eccentricities. The residual peripheral error may correspond to "opposite side" target misses cases (figure 4C), when the target is shifted away from the visual field horizon, and the agent can not recover from its initial error.

Concurrent action selection

Finally, when both pathways are assumed working in parallel, each one may be used concurrently to choose the most appropriate action. Two concurrent accuracies are indeed predicted through separate processing pathways, namely the central pixels recognition accuracy through the "What" pathway, and the log-polar accuracy map through the "Where" pathway. The central accuracy may thus be compared with the maximal accuracy as predicted by the "Where" pathway.

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:

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

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.

In our simulation results, the central accuracy is found to overtake the maximal peripheral accuracy when the target is close to the center of gaze. When closely inspecting the 1-10 pixels eccentricity range (not shown), a decision frontier between a positive and a negative information gain is found to lie at 2-3 pixels away from the center. Inside that range, no additional saccade is expected to be produced, and a categorical response should be given instead. While this frontier is not attained, micro-saccades may be pursued in the close vicinity of the target in search of a perfect centering. In the opposite case, when the central accuracy estimate is very poor, the comparison can still be considered helpful, for it may allow to "explain away" the current center of gaze and its neighborhood, encouraging to actuate long-range saccades toward less salient peripheral positions, making it easier to escape from initial prediction errors. 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" [**?**]. Combining accuracy predictions from each pathway may thus allow to refine saccades selection in a way that complies with biological vision. 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.

Quantitative role of parameters

In addition, we controlled that these results are robust to changes in an individual experimental or network parameters from the default parameters (see Figure 7). From

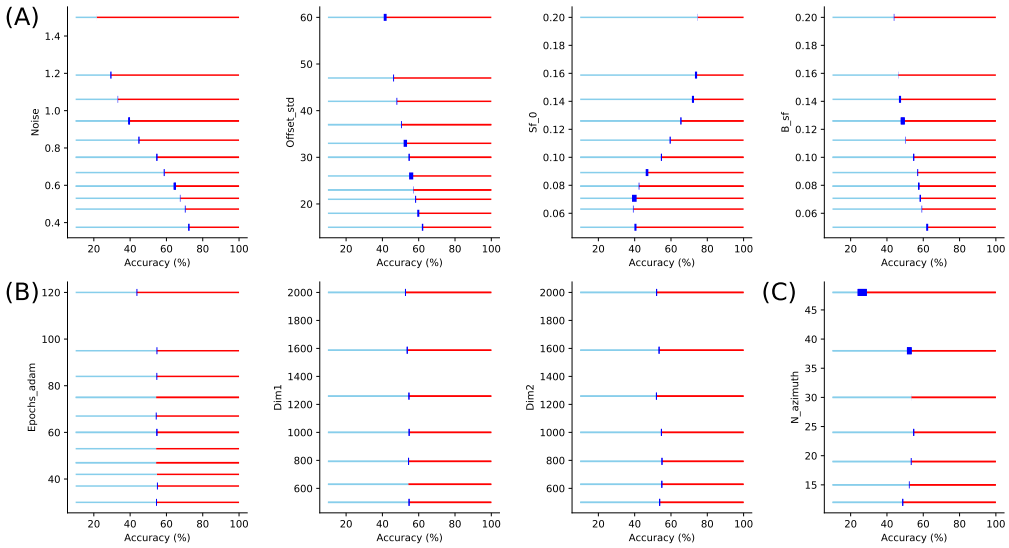

Fig 7. Quantitative role of parameters: We show here variations of the average accuracy as a function of some free parameters of the model. All parameters of the presented model were tested, from the architecture of image generation, to the parameters of the neural network implementing the "Where" pathway (including meta-parameters of the learning paradigm). We show here the results which show the most significative impact on average accuracy. **(A)** First, we 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 5 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. **(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 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 weak changes in accuracy.

the scan of each of these parameters, the following observations were remarkable. First we verified that accuracy decreased when `noise` increased and while the bandwidth of the noise imported weakly, the spatial frequency of the noise was an important factor. In particular, final accuracy was worst for `sf_0` ≈ 0.07, that is when the characteristic textures elements were close to the characteristic size of the objects. Second, we saw that the dimension of the "Where" network was optimal for a dimensionality similar to that of the input but that this mattered weakly. The dimensionality of the log-polar map is more important. The analysis proved that an optimal accuracy was achieved when using a number of 24 azimuthal directions. Indeed, a finer log-polar grid requires more epochs to converge and may result in an over-fitting phenomenon hindering the final accuracy. Such fine tuning of parameters may prove to be important in practical applications and to optimize the compromise between accuracy and compression.

Relation with other models

Our model is, to our best knowledge, the first case of a bio-realistic log-polar implementations of an active vision framework. We have thus provided a proof of concept that a log-polar encoding retina can efficiently serve object detection and identification over wide visual displays.

There are however lots of model that reflect to some degree the biological principles of sequential visual processing. First, active vision is of course an important topic in mainstream computer vision. In the case of image classification, it is considered as a way to improve object recognition by progressively increasing the definition over identified regions of interest, referred as "recurrent attention" [**?**, **?**]. Standing on a similar mathematical background, recurrent attention is however at odd with the functioning of biological systems, with a mere distant analogy with the retinal principles of foveal-surround visual definition.

Phenomenological bio-realistic models, such as the one proposed in Najemnik and Geisler's seminal paper [**?**], rely on a rough simplification, with foveal center-surround acuity modeled as a response curve. Despite providing a bio-realistic account of sequential visual search, the model owns no foveal image processing implementation. Stemming on Najemnik and Geisler's principles, a trainable center-surround processing system was proposed in [**?**], with a sequential scan of an image in a face-detection task, however the visual search task here relies on a systematic scan over degraded image, with visual processing delegated to standard feature detectors.

Denil at al's paper [**?**] is probably the one that shows the closest correspondence with our setup. It owns an identity pathway and a control pathway, in a What/Where fashion, just as ours. Interestingly, only the "what" pathway is neurally implemented using a random foveal/multi-fixation scan within the fixation zone. The "Where" pathway, in contrast, mainly implements object tracking, using particle filtering with a separately learned generative process. The direction of gaze is here chosen so as to minimize the target position, speed and scale uncertainty, using the variance of the future beliefs as an uncertainty metric. The control part is thus much similar to a dynamic ROI tracking algorithm, with no direct correspondence with foveal visual search, or with the capability to recognize the target.

Discussion

In summary, we have proposed a visuo-motor action-selection model that implements a focal accuracy-seeking policy across the image. Our main modeling assumption here is an *accuracy-driven* monitoring of action, stating in short that the ventral classification accuracy drives the dorsal selection on an accuracy map. 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 concurrent accuracy predictions (such as the one done through the "What" pathway) to bring out more elaborate decision making which are relevant for visual search, such as the inhibition of return [**?**]. This combination of a scalar drive with action selection is reminiscent of the actor/critic principle proposed for long time in the reinforcement learning community [**?**]. In biology, the ventral and the dorsolateral division of the striatum have been suggested to implement such an actor-critic separation [**?**, **?**]. Consistently with those findings, our central accuracy drive and peripheral action selection map can respectively be considered as the "critic" and the "actor" of an accuracy-driven action selection scheme, with foveal identification/desambiguation taken as a "visual reward".

Moreover, one crucial aspect of vision highlighted by our model is the importance of

centering objects in recognition. Despite the robust translation invariance observed on the "What" pathway, there is small radius of 2-3 pixels around the target's center that needs to be respected to maximize the classification accuracy. This relates to the idea of finding an absolute referential for an object, for which the recognition is easier. If the center of fixation is fixed, the log-polar encoding of an object has the notable properties to map object rotations and scalings toward translations in the radial and angular directions of the visual domain [**?**]. The translation invariance found in convolutional processing may thus be extended to both rotation and scale invariance in the log-polar domain. Incorporating this scale and rotation invariance may thus extend the generalization capabilities of the model.

Despite its simplicity, the generative model used to generate our visual display allowed to assess the effectiveness and robustness of our learning scheme, that should be extended to more complex displays and more realistic closed-loop setups. On the one side, the restricted 28×28 input used for the foveal processing is a mere placeholder, that should be replaced by more elaborate image processing frameworks, such as Inception [**?**] or VGG-19 [**?**], that can handle natural image classification. The main advantage of our peripheral image processing is its energy-efficiency. Our full log-polar processing pathway consistently conserves the high compression rate performed by retina and V1 encoding up to the action selection level. The organization of both the visual filters and the action maps in concentric log-polar elements, with radially exponentially growing spatial covering, can thus serve as a baseline for a future sub-linear (logarithmic) visual search in computer vision. This may allow to detect an object in large visual environments at little cost, which should be particularly beneficial when the computing resources are under constraint, such as for drones or mobile robots.

Finally, our model relies on a strong idealization, assuming the presence of a unique target. The presence of many targets in a scene should be addressed, which amounts to sequentially select targets, in combination with implementing an inhibition of return mechanism. This would generate more realistic visual scan-paths over images. Actual visual scan path over images could also be used to provide priors over action selection maps that should improve realism. Identified regions of interest may then be compared with the baseline bottom-up approaches, such as the low-level feature-based saliency maps [**?**]. Maximizing the Information Gain over multiple targets needs to be envisioned with a more refined probabilistic framework, including mutual exclusion over overt and covert targets. How the brain may combine and integrate these various probabilities is still an open question, that amounts to the fundamental binding problem.