Active Vision : on the relevance of a Bio-inspired approach for object detection

Kevin Hoang, Alexandre Pitti, Jean-François Goudou, Jean-Yves Dufour, Philippe Gaussier

Active Vision : on the relevance of a Bio-inspired approach for object detection

Kevin Hoang[1,2], Alexandre Pitti[1], Jean-Francois Goudou[2,3], Jean-Yves Dufour[2] and Philippe Gaussier[1]

[1]ETIS UMR 8051/ENSEA, University of Cergy-Pontoise, France, [2] Thales SIX GTS-Vision and Sensing laboratory, Palaiseau, France, [3] work realised during presence at Thales

E-mail: kevin.hoang@ensea.fr, alexandre.pitti@ensea.fr, jean-francois.goudou@thalesgroup.com, jean-yves.dufour@thalesgroup.com, philippe.gaussier@ensea.fr

Abstract.

Starting from biological systems, we review the interest of active perception for object recognition in an autonomous system. Foveated vision and control of the eye saccade introduce strong benefits related to the differentiation of a "what" pathway recognizing some local parts in the image and a "where" pathway related to moving the fovea in that part of the image. Experiments on a dataset illustrate the capability of our model to deal with complex visual scenes. The results enlighten the interest of top-down contextual information to serialize the exploration and to perform some kind of hypothesis test. Moreover learning to control the occular saccade from the previous one can help reducing the exploration area and improve the recognition performances. Yet our results show that the selection of the next saccade should take into account broader statistical information. This opens new avenues for the control of the ocular saccades and the active exploration of complex visual scenes.

1. Introduction

Understanding and interpreting our visual environment is a complex and energy-consuming task. Making the most of their physical resources, biological systems exploit different neural mechanisms to reduce the computational load [1, 2, 3, 4], in an attempt to make the rich input stream of data more manageable [5]. For instance, humans build their internal visual representation by integrating successive samples of the scene, focusing their attention on the parts carrying the most relevant information [6]. The eye itself is a testimony of this selective information, with the highest acuity located at the fovea and decreasing at the periphery of the retina.

As a result, a parsimonious exploration of the visual scene involves complex neural mechanisms to move the gaze with an eye saccade [7, 8, 9] in order to sample the visual

information efficiently on the fovea. In a relative minimum of time, this process allows to know *where* to look at and to predict *what* we should expect [10, 11].

Autonomous systems designed to operate in the real world face the same level of complexity. However, most of the state-of-the-art perception methods rely on powerful convolutional neural networks (CNN) which remain heavily based on the traditional sliding window paradigm [12, 13, 14, 15] and thus highly ressource-consuming. Even though reducing the CNN computational cost has been the subject of several recent investigations [16, 17, 18], all proposed methods still analyze the image exhaustively. As a result, CNN feature extractors [19, 20] tend to monopolize most of the computational resources, while their training notoriously requires a substantial amount of annotated data.

To reduce the computational load, several works [21, 22] highlighted the benefits of limiting the complex recognition process to regions of interest previously detected by a less costly operator. In a similar way, other methods [23, 24, 25] start by extracting points of interest in the image. Yet, those approaches compute the recognition features for all detected regions of interest [26, 27], resulting in a high computational cost when dealing with a complex scene. Furthermore, they assume that the visual information would be fully available at once, which is not always the case in particular for a perceptive agent evolving in an open world with a narrow field of view.

Going even further, other approaches draw inspiration from biological systems to propose attention-based strategies [28]. They consider the dynamic aspect of the perception process where an intelligent agent *actively* gathers samples from the environment where it carries the most relevant information . With the common goal of guiding a *virtual* fovea to the most important parts of the visual scene, they question whether a gaze movement control could be modeled from bottom-up low-level visual features (in a main saliency map [29, 30, 31]) or from top-down factors to optimize a predefined visual search task [32, 33, 34, 35].

Our work is inscribed in the active perception paradigm, tackling the visual object search problem from its computational constraint. We take inspiration from the way superior vertebrates are able to acquire the right amount of information by focusing their attention on the most relevant parts of their surroundings.

The remainder of this paper is structured as follows. In Section I, we show that a mobile eye with a logarithmic topology is an energy-saving sensor, which encourages to focus on a specific point of interest, like an object corner. In Section II, we conclude that with only a small part of the whole image processed, several samples should be analyzed for a robust recognition. We discuss how various exploration strategies should impact the performances and the resource consumption.

In section III, we present the Active Vision Architecture as an artificial neural network architecture to support our arguments. As illustrated in Figure 1, the proposed model processes sequentially small samples at various locations in the image. In this regard, it follows a conventional *feature extraction - feature grouping - object hypothesis/verification - object recognition* pipeline [36, 37, 38]. We model the first

Figure 1. General presentation of an active vision architecture perfoming an object recognition task. An iterative process extracts and store local features over time, while the captured information guides the exploration at each step. The system gradually accumulates evidence and emit hypotheses, until the confidence in the object identity is reached.

stages of the visual cortex to extract focus points and explore them sequentially. Then, this serial inspection performs two independent operations: the localization and the recognition of a target object. We model the object localization from the joint contribution of several neural fields. In parallel, the sequence of extracted local samples builds up a robust dynamic representation of the object for its categorization, formed by a merged sensori-motor information ("what" and "where") similar to the models proposed by [39, 40]. Finally, the same architecture is used to illustrate an example of learning a reflex saccade directly from the visual stimulus.

In section IV and V, we assess the viability of the model for an object recognition task. We evaluate its accuracy on a dataset of static images, on which we perform online incremental learning from only a few examples per object class to successfully localize and identify a target. More precisely, we will show the capabilities of a sensorimotor approach for object recognition, and we underline that the motor pathway of an active process provides a rich and more robust information, compared to a passive visual content. Then, we highlight the gain in computational cost compared to a state-of-the-art deep neural network (DNN). The chosen DNN is a typical example of a massively parallel processing architecture. With this comparison we highlight the core differences between an exhaustive and a parsimonious approach, and how it translates during the training and the inference. Finally, we evaluate the impact of a top-down saccade control strategy. We identify in which scenarios it actually succeeds or fails to improve the speed and accuracy of the model. The benefits and limitations of such a system open discussion to different strategies of saccade control in section VI.

2. The physical constraints of the visual system

We start by reviewing the intrinsic physical limitations of any agent endowed with visual perception. To detect the presence of a particular object, a first possible solution is to perform a parallel search of the target in the whole image. Thus, localizing an object would imply to duplicate the neurons sensitive to this object for each possible location in the image. In practice, it can be modeled by a convolution of the whole image with the different 2D views of the object that would represent all its possible variations of size and orientation, multiplying the number of convolutions accordingly. Therefore, dealing exhaustively with all the possible translations, scales, rotations would require a prohibitive‡ amount of computational resources [41], which would grow at least

‡ in a classical 640x480 image, for an object projected on a 40x40 pixels area in 2D, then 264000 neurons are necessary to detect its best correlation in the image taking into account the convolution

proportionally with the number of objects [36]. In this way, Tsotsos et al. first conclude to the need of a recognition from a partial visual input and a sequential attention mechanism. It turns out to be coherent with the anisotopic anatomy of the eye and the retina. Indeed, to make the connection with the biology, the latter can be seen both as a constraint and an advantage for an attention mechanism. On the one hand, it can be imagined as the evolution of a large high resolution visual sensor which preferred to concentrate the maximum number of photoreceptor in its center. On the other hand, the high number of receptors in the fovea gives a high resolution information, while the periphery gives a more global information sufficient to guide the focus [3]. In practice, the anatomy of the retina makes straightforward that in order to have access to a substantial amount of information, an agent can only accumulate samples from the whole visual input by guiding the fovea on several parts of the scene.

One way to model the discrepancy in the number of photoreceptors between the center of the retina and the periphery is to model the number of photoreceptors as a logarithmic function decreasing according to the eccentricity in the image, and more precisely by a log-polar distribution having its origin located at the center of the fovea [42, 43, 44].

Interesting properties of the log-polar distribution are exposed in appendix **??**. At a global level, those properties prove useful for object recognition [45, 46, **?**], in particular if the system can focus on the gravity center of the object as the origin of the log-polar referential. However, because the log-polar transform is highly dependent on the location of its origin, the system would become sensitive to occlusions or perturbations that would shift the position of the object gravity center. Hence, a more stable approach is to focus on peripheral corners and more generally on the end of lines.

Because an angle keeps its general shape after a rotation, a small skew or scale change, a focus point obtained by this method remains stable even after those geometrical transformations. In the primate visual cortex V1, receptive fields of cells from the primary cortex correspond to gabor-like functions that allow to detect oriented contour and end of lines. Thus in the following, to extract intersections as interest points, we model the primary visual area V1 (with a gradient computation and a convolution with a DoG filter).

From a more general point of view, focus points can be deduced from other image properties, as soon as they observe an abrupt change (visual motion, color). However, for the sake of simplicity and without any loss of generality, we will only consider the local gray-level contrast in static images for the rest of the paper.

On a side note, it is interesting to note that in the absence of corners, the focus points would not be stable with this method, making the object recognition impossible as observed in [47].

problems related to the borders

3. Performing an object recognition task

3.1. Scanpath : the visual perception as an exploration process

From the considerations of the previous section, we ask here how the visual information should be retrieve, as we now assume that an object recognition is achieved from a local view extracted around a given focus point. However, if we consider a complex object, then a single point might not be enough to recognize it [48], all the more if we take into account that the actual object area covered by a region of interest at a peripheral corner can be small. Consequently, an agent would need to gather additional information at several other locations in its environment. Now, if we suppose that it might have a limited number of visual sensors, then this recollection of visual information would have to be sequential.

Hence, we consider here the visual perception as a *dynamic* inspection of a scene. It leads to the idea of an agent that would analyze and further recognize an item by exploring sequentially several of its focus points, gradually accumulating evidence toward a threshold of decision. And for the following, the sequence of the explored local views will be referred to as a "scanpath".

The task is not trivial when the object is found in a complex scene. Indeed, in the presence of other items presenting potential focus points of their own, the system has to find the relevant points among an important number of less significant ones. The primate brain seems to have found a way of discriminating particular regions in its visual field as evidence of saliency has been found in the lateral intraparietal area (LIP) [49]. From a more generic point of view, two different approaches can be described. First, an exhaustive exploration of all the points surely insures to "see" at least once the desired object but can lead to a prohibitive processing time. On the opposite, a second solution would selectively prioritize the most relevant focus points (i.e. richer in information).

Considering a particular object as a target, it is an active strategy that optimizes the quantity of acquired data [50]. It would bring two benefits:

- a gain in processing time : if we assume that we know where to look to inspect a particular item, then the number of local views required for its recognition should be reached faster if we concentrate the focus only on it

- a gain in the quality of the recognition : the same way, looking only at the wanted target filters out any risk of false detection on information coming from the background or other foreign objects. Such a strategy requires a particular mechanism to favor its respective focus points, and later in the paper, we will tackle the subject with a learning approach.

A trade-off between those two representative strategies would be a random exploration of the image focus points.

Also, notice that robustness to occlusion implies a non-predefined scanpath, *i.e.* the freedom to explore the points in any particular order depending on the situation. Previous works [7, 8] corroborate those ideas, observing that when in search of an object the primate visual system moves its gaze from fixations to fixation, until it falls

approximately within the target [51].

A sequential exploration of the scene implies an inhibition of the already explored areas. This inhibition of return mechanism is modeled by a dynamical memory which stores the position of the previous saccades in an environmental referential [52, 53], more precisely object-centered [54].

For the following, we look into a 'spotlight' mechanism of selecting the visual information, regardless of whether it is done internally or by an actual physical movement of the eye. Furthermore, we will map those virtual saccades in the absolute referential, as we consider a static camera.

3.2. Integrating the information for the recognition

The question remains on how to make use of the information gathered at each exploration. Interestingly, two types of information are available after each ocular saccade, and previous works already established that separate cortical pathways are independently responsive to the content and the spatial information [10, 55]. After each saccade, the projection of the retina image on V1 provides a local view that can be recognized and named as the *what* information. At the same time, the position of the eye in the orbit or in the head referential is defined as the *where* information. We suppose the *what* to be computed in the inferotemporal cortex, while the control of the *where* is related to the parietal cortex. Following the idea of a sequential processing (see previously), it is straightforward to assume that we need a buffer to integrate the information gained at each local processing. Previous work [39, 40] observed that some areas in the hippocampus appear to be sensitive to both *what* and *where* information. Therefore, we tend to assume that a reliable representation of an object would be linked to a coherent combination between the *what* and *where*. Nevertheless in section 5.2, we question if a partial information integration could be enough in some cases to make an efficient object recognition.

4. Model and methods

In this section, we illustrate the above-mentioned mechanisms with the Active Vision Architecture (AVA) model, a neural architecture based on the work presented in [56]. This model has previously been exploited for navigation and object classification in a robotic context [57, 58, 59]. Here, we wish to utilize it for an object detection task in a set of complex static images. It is a general architecture which virtually reproduces the sequential analysis of local views, as explained in the previous chapter, and learns to performs three distinct operations (as depicted in Figure 2): the object localization, its recognition (or 'classification', as in the traditional computer vision vocabulary) and a proposition of the next saccade.

Beforehand, we train the model with several discrete views of the object, from which local views are extracted and stored as prototypes for further use (for more details on

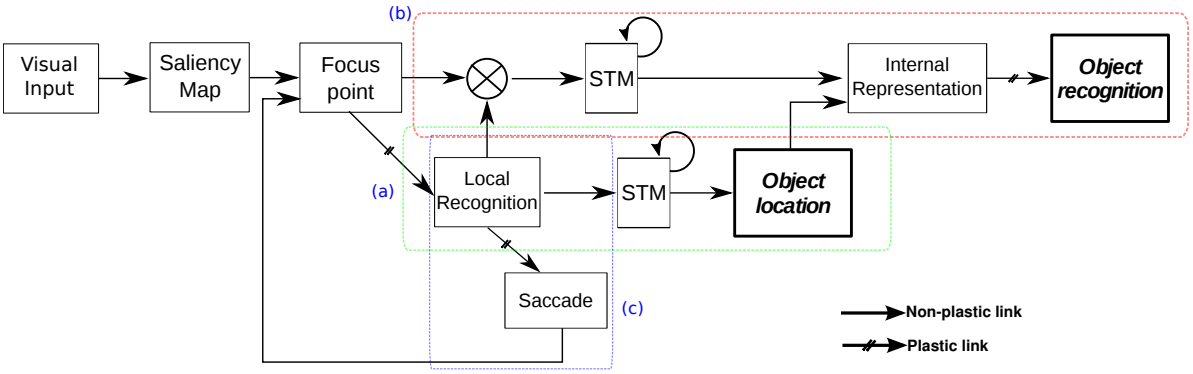

Figure 2. General architecture of the proposed system. The model performs three operations (a) The object localization (b) The object recognition (c) The control of the next saccade. Those three aspects are to be detailed respectively in 4.1, 4.2 and 4.4. The processing of the focus points being sequential, short-time-memories (STM) buffer the information over time.

how we characterize the latter, see section **??**). Afterwards, for a given focus point in a new observed scene, a similarity measure between the new local view and the learned prototypes provides a local cue about the object presence and identity in the image (Figure 2.a). Furthermore, rather than basing the recognition solely on a single sample from the input, the accumulated information over several focus points builds up an even more robust representation of the input to recognize (Figure 2.b). Finally, from the same extracted feature, the system learns to direct its gaze to another focus point belonging to the relevant object, regardless of any external perturbation (Figure 2.c). In this work, we use a general learning module combining an unsupervised phase of categorization which serves a dictionary of features for a second supervised learning phase (for more details, see **??**).

The following subsections detail the neural implementations and mechanisms performing the object localization, the object recognition and the next saccade proposition.

4.1. Object localization

To estimate the object position, we integrate the responses from neurons coding the relative distance of local views to the centroid (center) of the object. For better understanding, those neurons are organized in a way that respects the topography of the input image. If there is a spatial coherence in the found local views, the sum of responses coincides in a localized peak response. In addition, to insure robustness to small translation variations, we convolve the neural responses with a Gaussian kernel, equivalent to spread the neural activities over a neural field. Figure 3 illustrates this principle, with various level of recognition of a simple square-shaped object.

The AVA model learns to predict the object pose from the successive processed local views, and can be understood as an associative learning between their signature

vector and the object center position. Figure 4 illustrates the architecture for the object pose retrieval. For the sake of readability, only the y dimension is represented, while the model performs the operation in two dimensions.

First, the learning phase is realized with the object isolated in the image (a), to insure that all the learned local views actually belong to the object. Moreover, to free ourselves from the need of a groundtruth, the object is centered in the image. That way, learning the absolute position of a local view (i.e. distance from the zero of the allocentric referential) is equivalent to learn its relative distance to the object center.

The position of the currently explored focus points is coded in a neural 2D map, and a winner-take-all (WTA) mechanism combined with an inhibition-of-return process guarantees the sequential exploration of each point. Hence, during the learning phase: a neural group (Figure 4.b) codes the relative position of the object center from the current explored focus point. A neural group learning from the Least-Mean-Square (LMS) error (c) associates this position to the respective local view coded in a Self-Adaptive-Winner (SAW) neural group (d) from its signature vector (see details in section **??**). The section **??** details the definition, learning rules and equations for both LMS and SAW neurons. The SAW is a first stage of categorization which presents all the input local views as orthogonal vectors for the LMS.

During the test phase: as the position predicted by the LMS is relative to current focus point, we simply obtained the actual absolute position of the object center after shifting LMS value by the position of the currently explored focus point and we code it in a neural group (g). A Short Time Memory (STM) group (h) accumulates the LMS responses over all the focus points, and we simulate a neural field by summing the neural activities convoluted with a Gaussian kernel (not represented in the figure) for more robustness. Finally, a WTA competition (i) gives the position of the object center (j), as it should emerge as a coherent object.

During the training phase, the SAW learns a dictionary of categorizing features: it stores a local view prototype in each of its neurons, coding the input vector in its associated weights.

During the test phase, the SAW performs a similarity process : the local view signature is projected onto the space formed by the learned prototypes. After a WTA mechanism, only the neuron corresponding to the most similar feature is activated.

In addition to the location of the object, this object pose retrieval corresponds to the reconstruction of a visual referential [60] or 'zero' for the 'where' pathway, as a basis to build an invariant representation for the object recognition.

4.2. Internal object representation - Merging the 'What' and 'Where' informations

Figure 5 illustrates how the 'what-where' paths are merged in the AVA model for various local views extracted from an inspected object.

With the group (b) representing 'what' has been recognized and the group (c) representing 'where' it has been captured, a one-by-one product is performed (equivalent

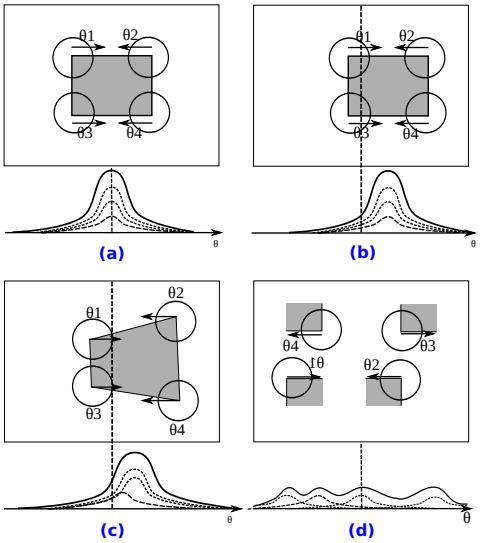

Figure 3. Principle of pose estimation with integration of neural activities: (a) four neurons learn to predict the center of a square object, when they encounter its four corners (b) when the square is translated the four predictions move accordingly, and the position of the activity accumulation predicts the center of the square position (c) the four corner are less recognized, but they still induce a peak of activity around the square center (d) when the object is incoherent, the sum of the activities is not high enough to highlight the position of the square.

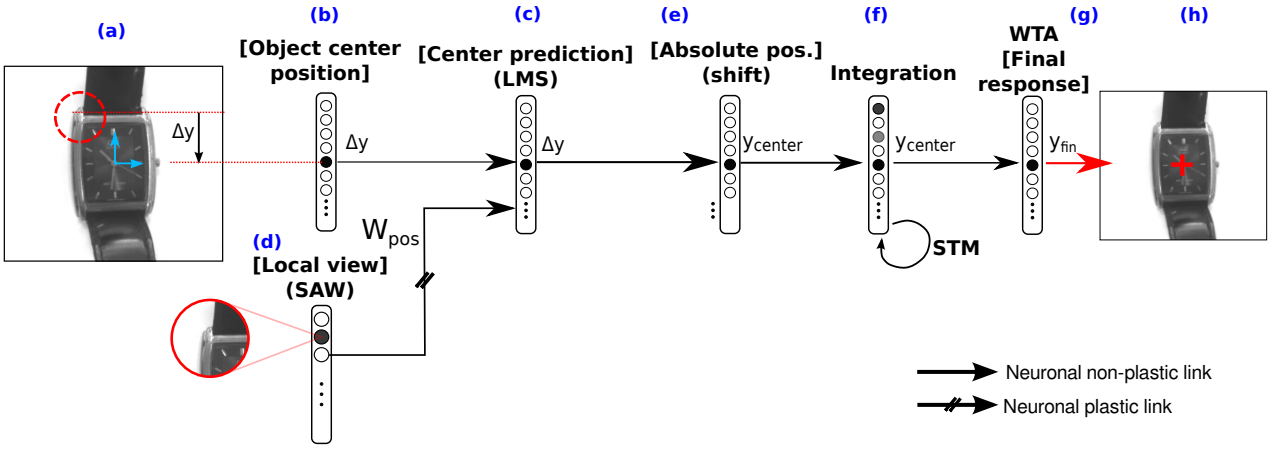

Figure 4. Localization principle with an integration architecture -detailed view of Figure 2-a : (a) a local view is analyzed on an object to localize (watch). During the learning phase: a neural group (b) codes the position of the object center and serves as a supervision signal for the Least-Mean-Square neural group (LMS) (c) which learns to predict the relative position of the object center from the respective local view coded in a Self-Adaptive-Winner neural group (SAW) (d). During the test phase: the group (c) predicts the relative position of the local view coded in (d) to the object center. The neural group (e) shifts it to code the absolute position in the image . A Short-Time-Memory (STM) (f) integrates the responses for all the explored focus points. A coherent object center emerges as a winner in a Winner-Take-All (WTA) group (g), and the object center of mass is localized (h)

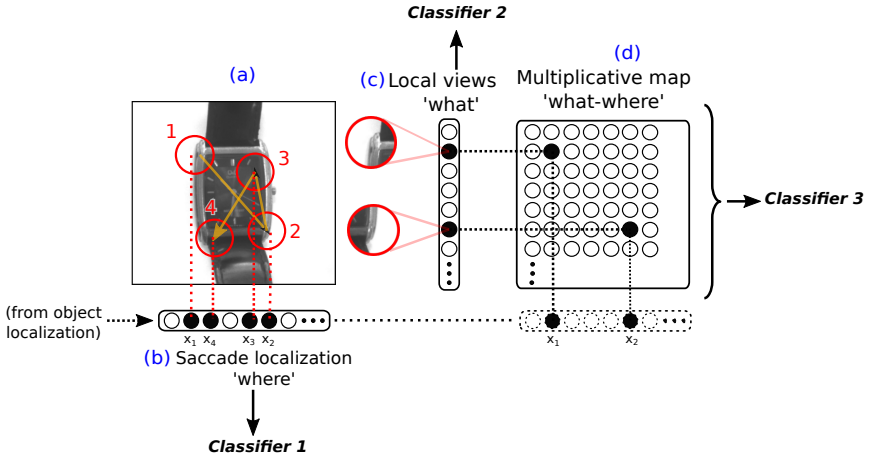

Figure 5. Internal object representation for recognition -detailed view of Figure 2-b. (a) illustrates an object (watch) and a sequence of four focus explored. A neural group (b) codes their respective x_i positions. (c) a neural group (SAW) codes the local view from their signature vectors. The neural group (d) represented as a 2D map illustrates a one-to-one multiplication between the first two local views processed and the position of their respective focus points, while a STM allows to integrate over the whole image exploration.
Section 5.2 evaluates if a system can recognize an object if relying solely on the positions of the focus points, the local views, or the fusion of both.

to an 'AND' logical operation), so we obtain a neural map (d) where a particular neuron is activated only when a specific local view is perceived at a given location, in the input image.

A short-time-memory mechanism maintains the activities of the neurons, integrating the information on N iterations, until a reset is triggered at the next visual scene to inspect. So the 'what-where' group, can be represented as a matrix of neurons M computed as:

$$M = \sum_{i=1}^{N} Rec_i.Loc_i^T \tag{1}$$

where Rec and Loc are respectively the vectors of recognitions and locations.

Over the exploration of N focus points, this group induces a pattern of neural activities which can be categorized by any classification mechanism.

4.3. Recursive feedback inhibition to deal with multiple objects

From the previous sections, the object localization (section 4.1) appears as a focus mechanism for the object recognition (section 4.2). As the system is strongly based on attention and competition mechanisms, dealing with multiple objects a the same time is intrinsically incompatible. So in order to detect multiple targets, we use a recursive inhibition process each time an object is successfully detected. Figure 6 illustrates how the object class recognition provides an inhibition feedback that filters out the contribution of redundant inputs. During the training phase, a neural group associates

the local views to the class of the object they belong to (Figure 6.b). That way, during the test phase, the system is able to assume which object it is analyzing. An one-to-one multiplicative operation (or 'AND' logical operation) with the last recognized class is then used to inhibit the local view at the localization process. This filtering process is the result of a point-to-point multiplication between two neural groups of the same size:

$$x_f = x_1 \times x_2 \tag{2}$$

In other words, if an incoming local view is assumed to belong to an already detected object, then it will not contribute to the localization of a new item. Consequently, the system should not focus on a previoulsy found target.

By extension, this recursive inhibition allows to cycle through all the learned object classes or can be used as a top-down factor to target specific objects.

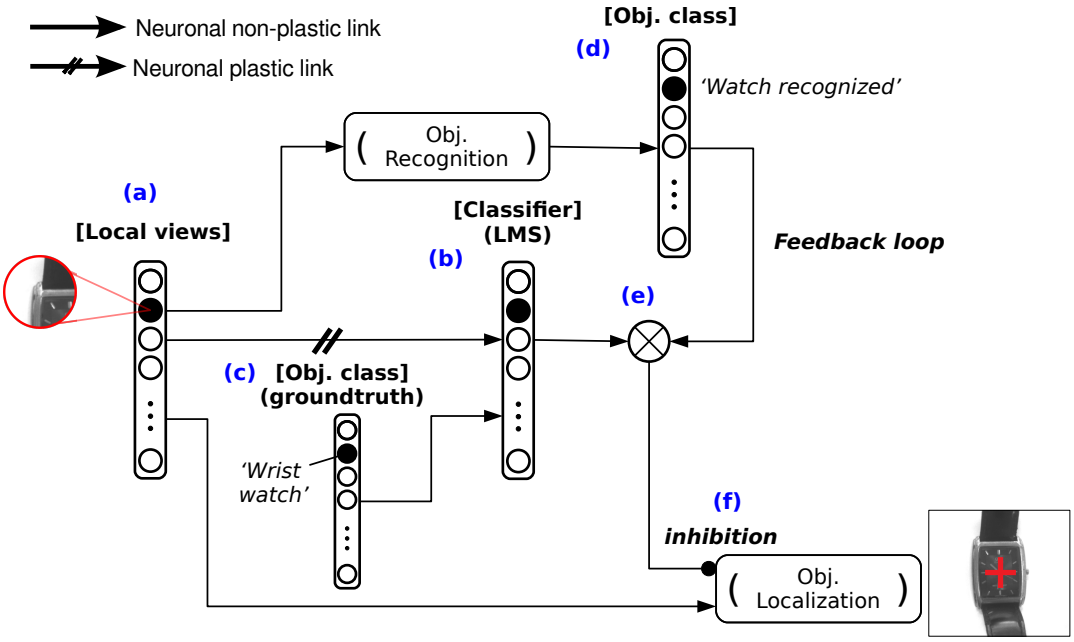

Figure 6. Feedback inhibition loop to manage multiple objects: During the learning phase, the neural group (b) associates the local views coded in the neural group (a) to the object they belong to, in a supervised fashion (c) (see **??** for the details). During the test phase : A multiplicative process (e) is performed between the predicted class in (b) and a feedback return from the object recognition (d). The result is an inhibition process for the Object localization (f). That way the system should not focus on a particular object if it has already been detected.

4.4. Learning the scanpath

With the benefits of an actively controlled scanpath already exposed in section 3.1, we propose below a strategy within the AVA model to illustrate our arguments. Because the system learns with an isolated object, we expect the scanpath to follow natural saliency highly different in the test image than the train image. Indeed, in a natural cluttered

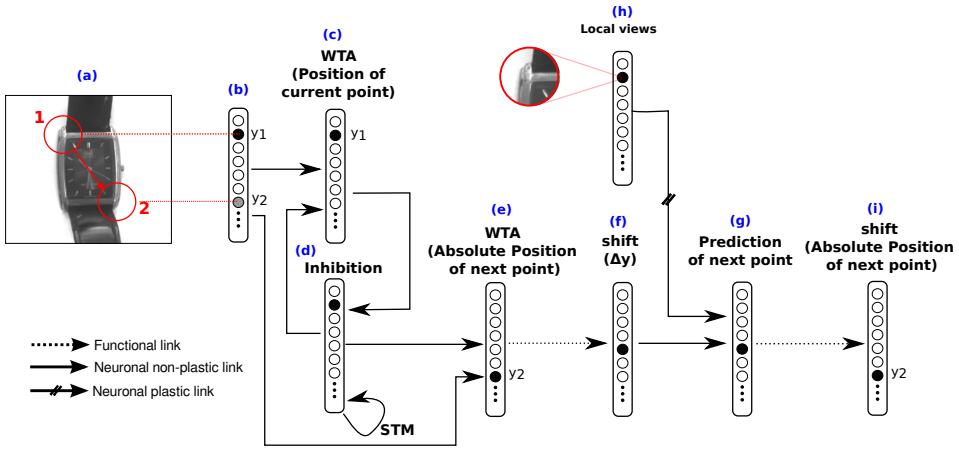

Figure 7. Learning of the next saccade -detailed view of the figure 2-c. (a) illustrates an object (watch) and a sequence of two focus points explored. A neural group (b) codes their respective y_i positions. (c) a WTA neural group position of the current focus point, from where a local view is processed. An inhibition of return (d) allows the sequential processing of the points, as the winner in the group (c) will be the second focus point at the next iteration. Though, taking in account the inhibition at the same iteration gives the position of the next focus point, coded in the group (e). During the learning phase, a LMS group (g) associates the relative position of the next point coded in (f) to the current local view coded in the SAW group (h). During the test phase, the absolute position of the next saccade is obtain by shifting (i) the position given by the LMS from a new input local view. This position can be used instead of the one provided by the (c) group to direct the gaze of the agent.

scene, the object to recognize will be present with a highly distracting background, saliency-wise.

Considering the saliency of the focus points *a priori* , we expect our system to randomly explore the entire image, following the intensity order of the points of interest (see Figure 7). So, our goal is to limit this exploration only to the targeted object to recognize. That way, we expect two consequences: first, we should improve the speed rate of the algorithm by reducing the number of focus points explored. Second, we should improve the recognition rate of the system by reducing the number of unwanted analyzed local views and thus the possibilities of mistakes.

In order to learn the scanpath, we propose the model as shown in Figure 7. For the sake of readability, only the y dimension is represented in the figure. Each focus point is coded in a neural map (c), which is the winner of a global competition between all the salient points in (b). In the figure, the group labeled (d) is an inhibition of return which insures the sequential exploration of all the points. Indeed, the focus point explored at the next iteration is the one whose saliency is the highest, after to the currently explored point.

Although, at the same iteration, an immediate inhibition gives the next focus point, coded in the group (e), while a simple translation represents the relative distance

between the current and the next focus point. Actually, the group (e) codes the saccade to make from the current point in order to reach the next one.

During the Training phase, similarly to what we utilizes in section 4.1, a local view is extracted around the current focus point and its signature is categorized by a SAW group, labeled (g) in the figure. Then an LMS supervised learning group (h) associates this local view to the saccade to reach the next focus point.

In short, we consider the scanpath followed during the training phase as optimal (when the object is presented isolated and centered). Then, we wish to reproduce it when the object is presented in a cluttered scene with other unknown items.

During the test phase, when the system recognizes a local view similar to what has been learned, the LMS gives the saccade to reach the point that have been explored during the training phase for the same local view. In Figure 7, the group (h), insures that the saccade is made respective to the position of the currently explored point. Finally, we use a WTA mechanism to regulate the balance between a learned and a random exploration. In practice, we would favor the learned point over the naturally salient one by increasing the weight of the connection coming from the group (i).

5. Experiments

In this section, we are conducting three evaluations:

(i) The comparison of performance when using three different types of representations for the object description(*where-only*, *what-only* or *what-where* fusion)

(ii) The computational complexity evaluation compared to a massively parallel system

(iii) The evaluation of the benefits and limits of a saccade control strategy

Those experiments have been carried out on a Linux platform with an Intel Core i5-6300HQ, with four cores cadenced at 3.6GHz, 16GB RAM. For the AVA model, the active exploration of an image according to a given context is performed at around 10Hz.

5.1. Dataset and performance measurement

We evaluate our model on a modified version of the dataset used in [61]. The latter is interesting for our application, as it presents isolated centered objects for training, as well as cluttered natural scenes for testing.

For the training dataset, we simplified the original one to keep only objects of similar dimensions, resulting in 110 640x480 shots equally divided in 5 classes : *Apple, Teddybear, Truck Toy, Salt* and *Vase*. The objects are taken from various angles segmented and presented in the center of a black background. The Test dataset consists of 41 cluttered shots containing multiple objects present in a natural scene (see Figure 8).

Because the AVA model outputs the predicted object center of mass, we consider a result a true positive if the predicted position falls within the ground-truth bounding

Figure 8. Dataset used for the evaluation of the model: (Left) Examples of the 5 object classes, shot from various angles, on a black background. (Right) Natural cluttered scenes used for Testing.

box and the correct object class is output. The same way, a result is considered a false positive if the predicted center falls into the ground-truth bounding box but the object is mistaken with another one, or if a prediction is out of any region delimited by a ground-truth. We calculate the recall and precision from the true/false positives and the ground-truth and finally, evaluate the global accuracy performance of the system with the F_1 score (or f-score) [62] defined as follows:

$$F_1 = 2.\frac{precision.recall}{precision + recall} \tag{3}$$

5.2. Evaluation of the what-where robustness

The purpose of the following experiment is to highlight the robustness of a sensorimotor approach when recognizing an object class. As stated previously, we consider that the object recognition is performed through the accumulation of the joint information of *where* we looked and *what* we saw. However, we question here whether a simpler representation would be enough to successfully recognize a target. For that we compare three types of representation for a given object: during the visual exploration, we accumulate:

- the 'where' information (the locations of the saccades)
- the 'what' information (the visual content at the fixation)
- both above informations, merged into a 'what-where' tensor as in section 4.2

As illustrated in Figure 5, each representation is evaluated independently by a standard supervised classifier. In our case, we used the same learning module as in section **??** this time to train specific neurons (or "view cells") to respond to a given object category.

To help the system generalize, we artificially generated some new data (i.e data augmentation) to take into account variations in scale. We created additional images with ratios of 0.7 and 1.3 of their original values, which lead to a total of around 300 images for training. From those images, we set our system to learn in one-shot each local view, and store each one of them on one neuron of the first model stage. We define our *virtual fovea* as a circular region of 40 pixels radius. With around 10 fixations per training image, we code the dictionary of local views on 2500 neurons. A 2D map of 220x160 neurons predicts the object localization, following the process described in section 4.1. For the object recognition, we encode the train images (i.e. object views) on 300 neurons and classify them with 5 neurons representing the different classes. During the test phase, we set the tolerance of our recognition process low enough to generalize, in other terms to extrapolate to the new data.

The image exploration should interrupt if :
- it reaches an arbitrary upper limit of 50 saccades, which we estimated should allow the agent to cover the whole image in the majority of cases.
- the recognition confidence would reach an arbitrary threshold that would make the system successfully recognize an object.

Figure 9 shows the recall, precision and the accuracy for all three representations.

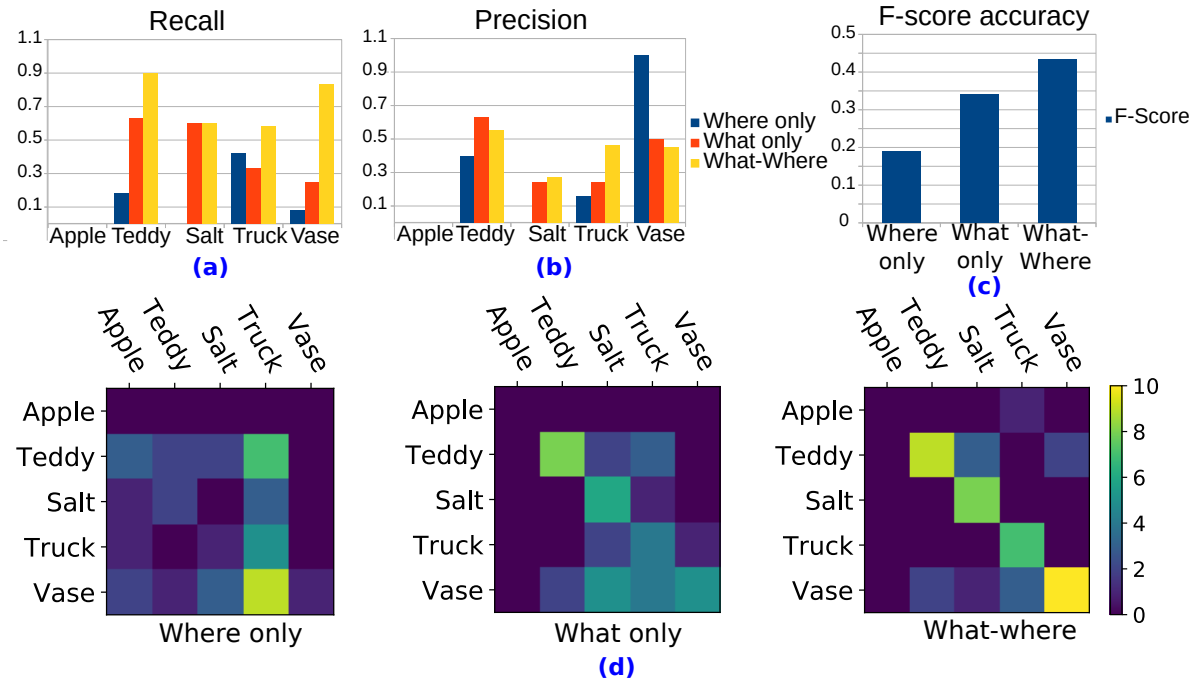

Figure 9. Evaluation of the *where*, *what* and *what-where* representation. (a) and (b) Precision and Recall measured for each five object classes. (c) Average accuracy f-score for three different internal representations. Merging the *what* and *where* informations provides the best results. (d) Confusion matrices for each internal representations. Relying on the *where*-only or *what*-only leads to confusions between object classes. The *what-where* representation removes ambiguities and results in a better accuracy.

First of all, it is worth noticing that none of the representations succeed to detect

the first object (*apple*). It is mainly due to the absence of stable focus points on this particular object and specific texture which make it difficult to detect (lack of corners, uniform texture). Thus, the comparative results will be based on the other objects.

The lowest accuracy is obtained by relying only on *where* the information is retrieved (i.e. location of the saccade). For an active perception system, it is actually possible to correctly detect an object using the motor path as the only information (as showed by the non-null Recall and Precision rate, Figure 9.a-b). However, this criteria is not differentiating enough between classes and the objects are often mixed up (see confusion matrix, Figure 9.d).

We should keep in mind that we are dealing with an object *detection* problem. Indeed, the system can miss an object present in the scene and does not always give an answer. So in the case of a non detection, the object is neither a True nor a False positive, and it won't appear in the confusion matrix.

A 'bag-of-words'-like approach proves effective in limited cases. In particular, because all the features are collected globally in the image, the system fails to deal with the ambiguity brought by multiple objects in the scene. Object like the *teddybear* and *vase* offer rich textures and details that prove useful for the detection, but also are a source of confusion when a local view may belong to several class (see confusion matrix Figure 9.d).

The sensorimotor representation (*what-where*) proves to be the most accurate for the object recognition (Figure 9.c). It removes the ambiguities encountered when dealing with only one type of information and helps for the object localization (the respective recall is the highest in this experiment, Figure 9.a) with less confusion between classes.

5.3. Comparison with an parallel processing system (DNN)

For the following evaluation we want to highlight the capabilities of a bio inspired process such as the AVA model compared to a massively parallel computing system. We will assess its detection effectiveness in terms of accuracy and its computational requirements both in training and inference.

By extension, we believe those aspects should mirror the suitability of a particular method for an autonomous system in an open world.

We compare here the AVA model to the Single Shot Detection model (SSD) proposed by Liu et al. [13] which is a state-of-the-art algorithm for object detection. Like similar other architectures [14, 15, 63], it is a Deep Neural Network method which performs a simultaneous localization and identification of the object from a forward pass after several stages of convolutional layers. Those models consider the object location extraction as a regression task, where convolutional layers learn to predict multiple bounding boxes that fit an observed object from a predetermined bounding box, called *prior* in the SSD model. An object is detected as far as the confidence of detection exceeds an arbitrary threshold set in a softmax layer (=0.6 in the original implementation). Then a non-maximum suppression operation keeps the

Figure 10. Images used to train the SSD model. The training examples have been generated from the segmented object inserted into random natural backgrounds. In addition, we resort to data augmentation, that presents the object rotated, scaled or cropped in the scene.

most prominent response as the final answer.

5.3.1. Training

Such a small dataset as in 5.1 is particularly a challenge for Deep network-like models, well known to learn on a substantial amount of data. Seemingly, an efficient training implies providing enough variety in the learned examples to avoid overfitting.

Because the training dataset always presents the examples on the same uniform background, we artificially generated 14000 images from the original 110 segmented training object by inserting them into a set of 40 random natural scenes and then applying some data augmentation where each object has been cropped, rotated and scaled (see section **??** for the parameters of the data augmentation). Considering the implicit additional data augmentation that comes with the original implementation of the SSD framework [64], we can estimate that a total of around 45000 images where used to train the model offline. Figure 10 illustrates some examples of resulting images. We only trained the specific SSD layers (transfer learning), as the implementation we used was based on a VGG-16 backbone [19] pre-trained on Imagenet-COCO [65].

5.3.2. Inference

When evaluating both the AVA and SSD models on the dataset, we obtain the results illustrated in Figure 11. The AVA model is more successful in finding an object, when each class is considered individually (see Recall rate Figure 11.a) However, the SSD model outperforms the AVA model in terms of overall F-score accuracy (Figure 11.d).

As stated in the previous evaluation, the AVA performances are intrinsically linked to the extraction of point of interest. Indeed, it appears that the objects presenting the

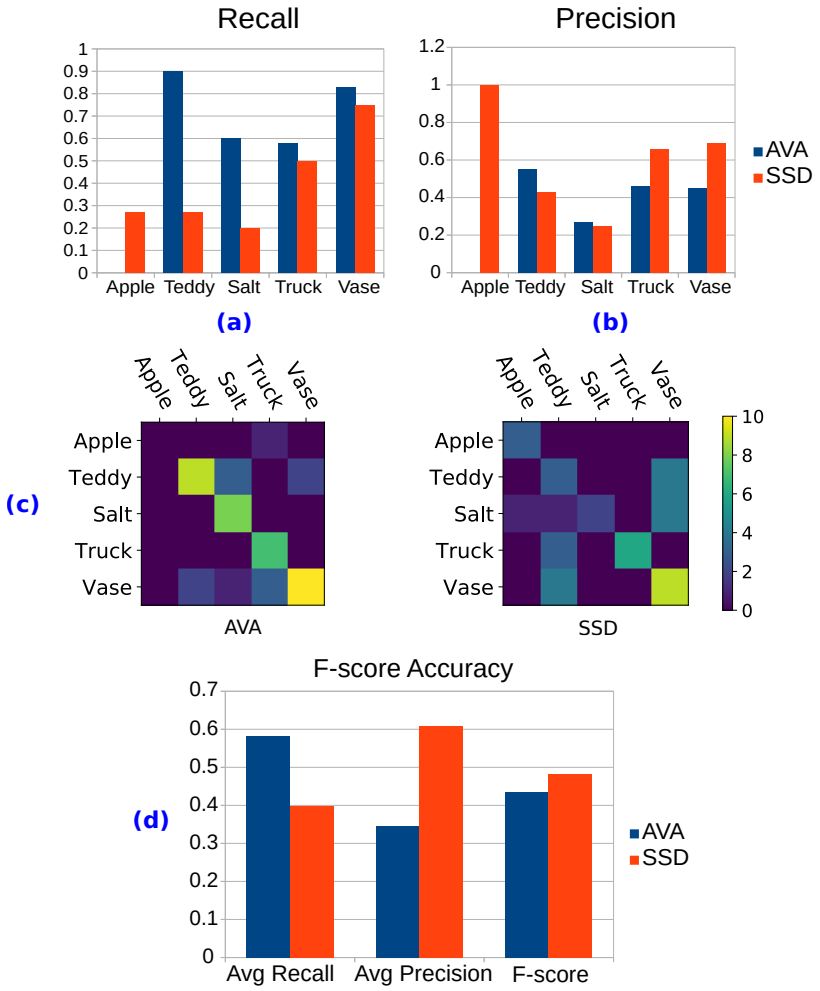

Figure 11. Accuracy performance test of vs SSD. (a) and (b) Precision and Recall measured for each five object classes. The AVA model seems more likely to detect the 'objectness' of the items while the SSD model performs a better recognition. (c) Confusion matrices : The SSD model detects at least once each object, but tend to be more subject to confusion betwen classes. (d) Overall, the SSD model outperforms the AVA model which obtains an accuracy of 0.44 compared to 0.48 for SSD

	MAC operations	Train examples	Epochs
SSD model	$31*10^9$	45000	3
AVA model	$1{,}02*10^9$ (average)	300	1 (one-shot)

Table 1. Comparison of the computational complexity. By design, the AVA model does not need to compute the same amount of data per frame. Considering a successful detection using around 10 saccades in average, we estimate the above mentioned number of MAC operations. We observe the same discrepancy for the training phase, with a substantial difference in the number of required training examples. The SSD model has been trained until convergence.

richest textures (*teddybear* and *vase*) are the most easily found. On the opposite, the system fails to detect the *apple*, mostly due to the lack of sharp angles and its quite uniform texture.

On the other hand, the SSD model succeeds to find at least once each object, displaying robustness to position, rotation, scale variation, and occlusion. The Precision rate (Figure 11.b) tends to emphasize the better recognition capabilities of the DNN. Intuitively, the multiple feature maps extracted by the CNN offer a richer description of the object than the *what-where* representation used here.

In our tests, we observed that the generation of artificial images has a major impact on the SSD ability to detect the presence of an object in the image (also described as "objectness" in [66]). Hence, the purpose of having numerous training examples is to provide enough variety and prevent the learning process from overfitting on a particular background. The data augmentation plays a significant part in the class recognition itself, once an object is pinpointed. Naturally, a particular type of data augmentation provides the system with some robustness to the respective transformation (for instance, zooming in/out on an object in the original training images allows the system to be robust to scale changes during the inference).

By design, the AVA model doesn't process as much data during the inference. As such, it doesn't need as much variety in its learning examples (for instance, the model intrinsically deals with possible occlusions as it only processes small parts from the object). As summarized in the table 1, this results in a substantial size difference of the learning data. For this particular experiments, we observe a 150 size ratio between the learning datasets used by both methods, and the AVA internal mechanisms allow to learn in one shot only.

We observe the same discrepancy in computational complexity. We evaluate it by measuring the number of multiply-accumulate operations (MAC) realized after detecting an object. This quantity remains constant for the SSD architecture as it computes each input image the same way. For the AVA model, the number of operations varies with the complexity of the scene (i.e. the available focus points, whether they cover an interesting region, the saliency order). Nevertheless, we observed an average of 10 saccades needed to detect an object with the AVA model, and we measured that it computes around 5% of MAC operations compared to its counterpart (see table 1).

The latter underlines how both methods differ in processing the data. In reality, the DNN-based model massively computes every location of the image in parallel and is able to detect multiple object at the end of a single forward pass. The AVA model only processes small parts of the scene and focuses on a particular target (see Figure 6 that shows how the system focuses on one particular target, or recursively on multiple objects). This makes it more suitable for *point-and-reach* -like operations or general applications where an actual target is defined. Also, it assumes that the agent can incrementally gather information that is not available at the moment. The confusion matrix in Figure 11.c also underlines how the AVA model tends to be more conservative than the SSD model, which achieves a better overall accuracy but is more subject to

confusion between object classes.

5.4. Evaluation of a Top-down saccade control

The purpose of the following experiment is to show how a saccade control strategy can optimize the scanpath and help towards a more effective recognition. (as discussed in 3.1 and 4.4). In the previous experiments with the AVA model, the system explored the focus points in a semi-random fashion following a bottom-up based visual saliency. As a result, it may have also processed unnecessary focus points that were not helpful in recognizing the object (typically points belonging to the background). We believe a correct object recognition is possible with a parsimonious number of saccade, so that actually computing the whole set of focus points should not be mandatory [26, 27].

A similar work has been carried out by Bicanski et al. [67] modeling ocular movements from one-shot learning mechanisms with grid cells. They obtained interesting results where the ocular scanpath help recognize an image within a few saccades. However, their results did not express the benefits of a controlled scanpath when generalizing to new data, as they evaluated their system on the learning examples.

We will now expose how the system behaves on various sets of data beginning with a simple toy example. Then we will evaluate the generalization capabilities by testing its behavior with complex images. In order to expose the effect of a controlled saccade we purposely favored the proposed location to guide the gaze over the random exploration by weighing accordingly both streams in a WTA mechanism.

First, we train and test the system on 640x480 images captured in real-time from a USB camera. The systems learns an object view and its respective scanpath from a single shot (here, a *pen* in Figure 13.a). To test the recognition, the objects are presented on a uniform background and moved around by hand without too much variation. That way, an new object will differ too much from a learning example (besides inherent small changes in illumination, rotation, scale). Nevertheless this simple setup illustrates how the gaze focuses on the target object (the pen), as well as other foreign items (packages, chewing-gum) (see Figure 13.b).

Figure 13.c illustrates the system behavior when it follows the learned scanpath. Once the simulated fovea inspects a part that belongs to a known object (the pen), the saccade control system guides the gaze from a local view to the next and the scanpath remains on the target. Thus, focusing on a single object allows to reduce the total number of explored focus point, speeding up at the same time the processing of the image.

The following shows how the system behaves on a set of complex images from the dataset (5.1) . Figure 14 illustrates the evolution of the recognition rate of the five different objects with each saccade. The individual object recognition is described by the activities of the 'what' neuron (see Figure 5) accumulated during the exploration of the whole image. This information translates into the current system confidence to recognize an object. As the agent gaze process the image randomly, we observed

some 'plateau' in the activity evolution of the recognition neurons. In absolute terms, an optimal recognition for a given object would translate into a steady increase of its respective neuron. For instance, in Figure 14.a, the consecutive analysis of the focus points 3, 10, 18 and 39 would make the neuron activity build up until the recognition threshold without interruption. Indeed in Figure 14.b, the cascade control strategy makes the system explore those focus points earlier, resulting in a faster recognition (in 25 cascades instead of 39). In addition, with a controlled scanpath the system naturally avoid some parts of the image that are a source of confusion (in Figure 14.a, the 35th saccade actually mislead the system toward the presence of an *apple*).

However, Figure 12 summarizes how the proposed strategy to control the scanpath fails to improve the performances at the global level. With an exploration following the bottom-up saliency of the image, the average number of performed saccades revolves around 20 before succeeding to find its target. The scanpath control actually results in longer explorations for some images of the dataset. As a result, the average number of saccades performed for the whole test set seems globally unaffected. It appears that the system does not have the ability to react accordingly if a misinterpretation is somehow introduced at the beginning of a scan path, as shown in Figure 15, and this local mistake is transmitted to the rest of the scanpath. With the gaze guided to erroneous areas, the system is more subject to perform a wrong recognition at the end of the exploration, accumulating local views from the background or foreign items. Indeed, small mistakes have been induced during the whole evaluation and the global f_1 accuracy performances are actually lower with the proposed particular saccade control.

6. Discussion and directions forward

This paper focuses on active vision, tackling the problem from the computational resource constraints. In the first experience, we exploit a random exploration strategy of the focus points as a good trivial heuristic between an exhaustive process and a very object-selective inspection, With that strategy, the AVA architecture illustrates how a general active model could be effective to detect multiple objects in a complex visual scene. It reveals resource-efficient by only exploiting a small part the original image for the object recognition (sampling an average of only 10% of the original input), and around 1% of the architecture is dedicated to the visual feature extraction process (represented by a convolution of four Gabor filters).

This small amount of resources allocated to characterize the visual input allows training the system in an online incremental fashion. In addition, learning from small local views makes the model all the more suited to deal with occlusions as far as the visible part of the object have been available during the training phase (see Figure 16). Hence, we underline a first advantage of integrating a 'motor' pathway in a recognition model for an active recollection of data: it increases the capabilities of the perceptual system whether it is for reducing the computational load with attention mechanisms [34, 35], or to have access to additional data that would remove any ambiguities for

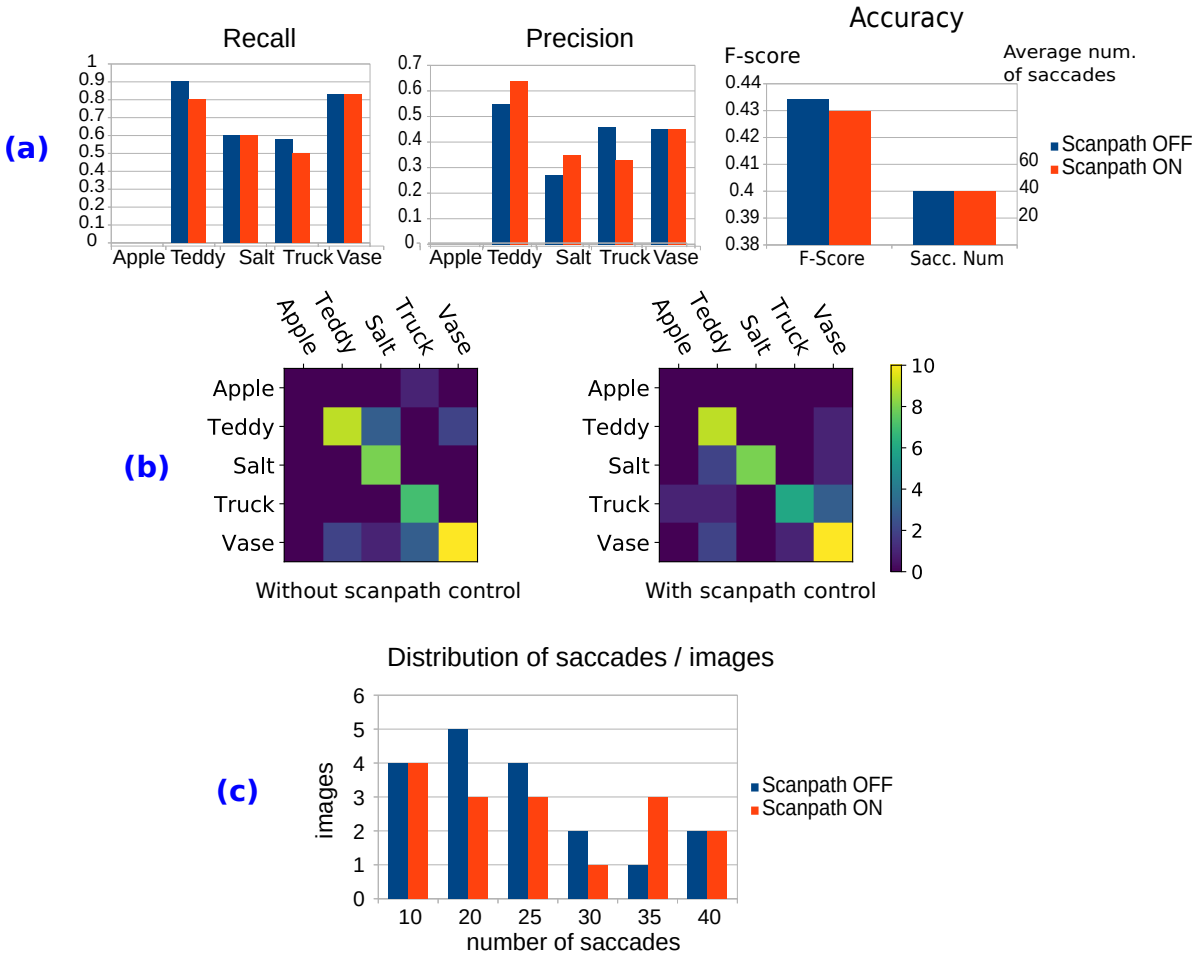

Figure 12. Accuracy performance evaluation of the AVA model, with/without saccade control (learned scanpath). (a) Precision, Recall and f1-score accuracy between the system with a random exploration of the focus points (blue) or with a saccade control (red). The proposed strategy globally does not have effect on the average number of saccade performed, while it seems to impact negatively the recognition performances. (b) Confusion matrices: the learned scanpath does not seem to improve the precision, and induces false detection that did not occur without saccade control. (c) for a random exploration, the number of saccades required for a correct object recognition revolves around 20 saccades (only the case where the object has been found before the maximum number of saccades is represented here). With the scanpath control, the system actually performs more saccades on particular images, thus failing to improve the global performances.

the recognition process [68, 69]. Secondly, the saccades provide a spatial information ('where') whose incorporation in the object representation proves effective for a robust recognition. Here, we use the 'where' information naturally provided by the 'motor' pathway to build a coherent representation of the object from raw visual perceptions as 'what' that an exhaustive approach tend to loose [70]. This representation proves more accurate than a "bag-of-feature"- like approach, robust to noisy and dense environment, where the presence of local views does not necessarily mean a coherent object (see Figure 17). This notion of spatial consistency, goes within similar lines with [71, 72] that also

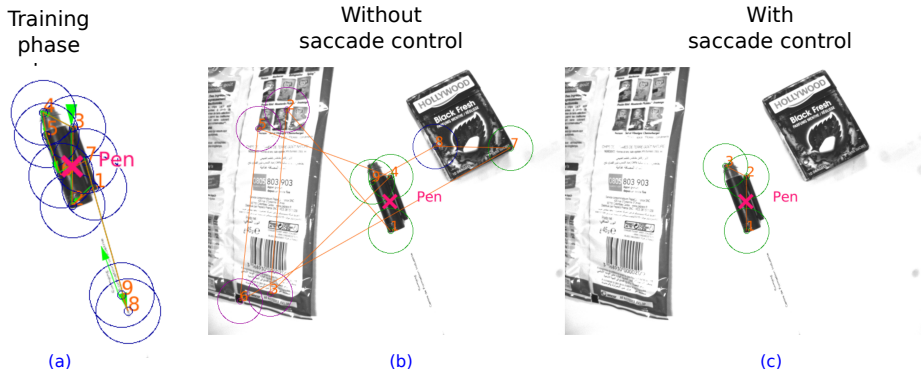

Figure 13. Toy example illustrating the proposed control of saccades. (a) An object ("pen") is learned from 9 saccades. The blue circles represent the local views processed by the "virtual" fovea exploring the image and the attached numbers shows the sequential order of the scan path. The light green arrows represent the learned scanpath. Then, the same object is to be found in another image in the presence of various other 'disruptive' items (chewing gum, packages,...) (b) With a random exploration the system explores a few focus points belonging to the plastic package, and the pack of chewing-gum. In fact, three saccades on the actual object (N.1-4-9) lead to a correct detection (c) With the saccade control, once a first focus point on the actual object is processed, the following scanpath remains on it and lead to the same number of processed local views for a correct recognition.

underlined the importance of a relationship between higher-level features.

In short, we simply believe that even if the primary motivation of an active system is an economical strategy, it should be also natural to make use of the available additional information for understanding what it sees.

However, this technique requires to explore the image long enough to actually *see* the target at some point. Also, we duly note that the evident sequential aspect of this approach implies the impossibility to parallelize the process. Somehow, a trade-off has to be found as the advantages brought by the active vision must compensate for the lost time. Thus in a second phase, we introduced a possible strategy to infer a scanpath, in an attempt to perform the object detection in a more controlled and effective way. The method proposed here is the direct associative learning from a bottom-up factor (a local low-level visual stimulus, represented in the local view). As such, it is highly dependent on its local perception.

Based on these observations, we review below some thoughts about the limitations of our architecture and reflect on ways to improve its performances.

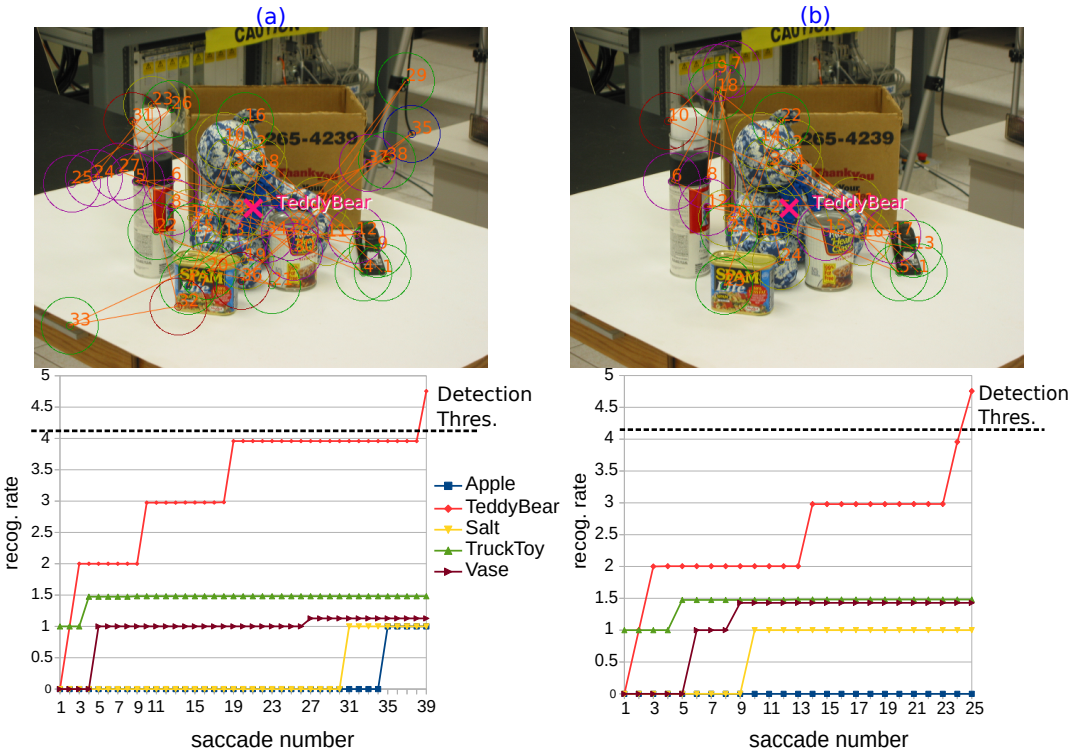

Figure 14. illustration of the proposed saccade control on the Office dataset. (a) The system randomly explores the scene and four saccades have been actually useful for the correct recognition of the *teddybear* object (N. 3-10-18 and 39), while a total of 40 saccades have been necessary (b) with a the saccade control, the same useful saccades are performed earlier (respectively here at N. 3-3-14 and 24) leading to only 25 saccades performed in total. However, while the *teddybear* has been correctly recognized, it is worth noticing that the *vase* obtained a better recognition rate in the second case, implying an exposition to false detections.

(i) *Improve the model recognition capabilities*

The accent put on the *active* aspect of the inference throughout this work should not diminish the importance of the local recognition from informative enough visual features. Hence, a perfect agent should be able to quickly recognize its target from a single glimpse, in particular for time-consuming considerations (see above). Nevertheless, we simply prefer to consider a pessimistic scenario where the single-shot recognition fails.

That being said, we do acknowledge the contribution of a more effective characterization of the local view for a better global object recognition [26, 25, 73].

(ii) *Extend the range of possible actions*

The general approach we presented here does not free itself from the paradigm of learning an object from single views [74, 37, 36] and in this method, it is very difficult to generalize over transformations or views inexistent during the training phase without relying on some artificial data augmentation. This property is also found with other *passive/exhaustive* approaches that would have no other choice than to

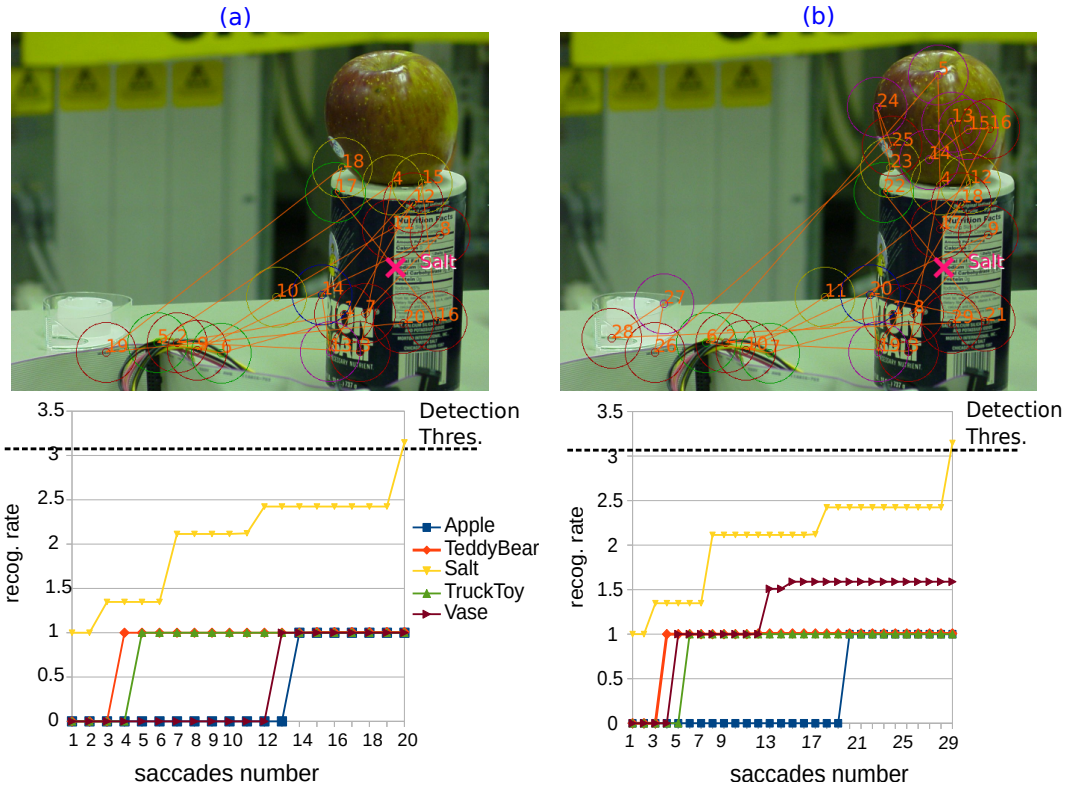

Figure 15. illustration of the proposed saccade control the negative impact on the Office dataset. (a) The system randomly explores the scene and four saccades have been actually useful for the correct recognition of the *salt* object (N. 3-7-16 and 20), while a total of 20 saccades have been necessary (b) with a the saccade control, the same useful saccades are performed later (respectively here at N. 3-8-21 and 29) leading to 30 saccades performed in total. It appears the system mistakes from saccade 13 and follows a wrong scanpath in order to find a wrong object (*vase*).

Figure 16. The AVA model succeeds to detect partially visible objects

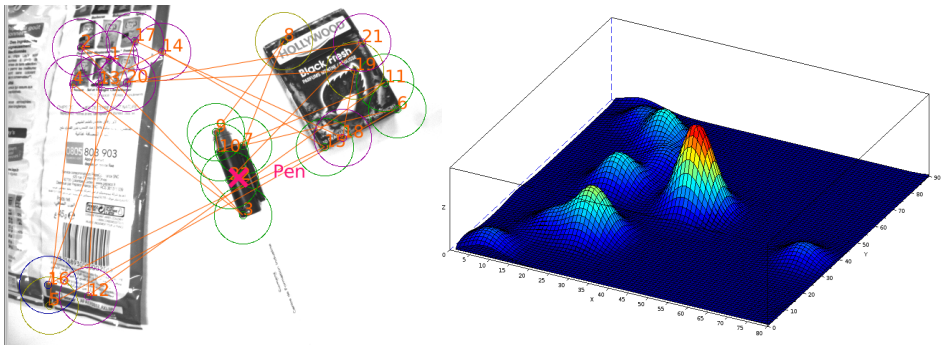

Figure 17. a) Visual exploration of an image to find the object in the center (a "pen") b) 2-D map of the neural responses responsible of the object localization. While each processed local view generates a localization prediction, the spatial coherence of the 'pen' integrated responses induces a stronger activity that wins the competition for the object localization.

go through the whole space of possibilities during training [75]. By comparison, a deep neural network [13], obtained similar results on the dataset used in this work, with a data augmentation that multiplied by 70 the original number of learning examples and with a rather *large* dictionary of visual feature (5330 filters). An improvement in the paradigm of *active* perception, should be to extend the range of the possible actions of the system, here only restricted to the data recollection process. Internal mechanisms could help the matching process between the new example and the learned prototype for a better identification [76, 77]. More than generalizing, the system would be able to recognize the transformation itself.

(iii) *Regulate the training phase*

In our application case, we forced our system to learn from the object view in one shot, with one neuron dedicated to code its respective local view, thus memorized as a prototype in our architecture for a further matching-like operation. It is coherent with a robotic application in an open world, where the agent has limited time to learn online from a single example. While we lower the recognition tolerance as a way of regularization, this approach however is inherent with over fitting to that single example. But learning is more than a direct association between whatever elements through memory and we are obviously looking for prediction to generalize to situations beyond particular examples previously observed in the past [78]. A first solution to lighten the strong relation between the recognition process and a learned prototype would be to avoid single-shot learning, for the system to spread the recognition over several training examples.

The same way, a second solution would be to spread the activity of the recognition over a sparse population code [79, 80].

(iv) *Countervailing the determinism of a reflex action*

Simply put, the direct associative learning presented here for the control of the next

saccade is similar to a conditioning from a given visual stimulus. Therefore, the decision of moving the gaze to the next location is solely dependent of the current local view and such a deterministic mechanism can induce errors when the agent is subject to a false local recognition, as seen in Figure 14 and 15. We consider that insuring a correct interpretation of the local visual information 100% of the time is likely to be unachievable in a complex real-world scenario. In the situation of an agent starting from an agnostic state, its first glimpse is primordial, as it defines the basis of the expectation for the rest of the exploration, thus defining a strong bias for the initialization of the scanpath. But a bias implies confidence, which, if misplaced, will induce negative repercussion for the rest of the scanpath by leading the gaze to wrong areas of the image and affect the global accuracy performance of the system. Hence, the inevitable case of a local mistake in the recognition process shows the local view and the decision process should be somehow independent [81].
It is important to notice that animals rarely find themselves without a contextual task. Their suppositions *a priori* are useful to lead its decision and more particularly its saccade movements [82]. The system should be driven by the task or a high-level belief of an expected object. This is consistent with the assumption of a target item considered as a top-down factor in a predictive approach [83], where the local views are collected in order to reinforce this assumption by maximizing the recognition of the object [32, 84] or reducing an entropy-like measure [33, 85].
From the existing AVA model, the feedback loop for object selection should be used as a feedback bias representing the target item to seek, or any other top-down motivation. Furthermore, even in the case of a correct local assumption, a deterministic saccade is not robust to occlusions, so the system should consider a broader range of possible location. Rather than guiding the gaze to a single point, the saccade control should increase the attractiveness of a larger area, modeling a 2D map of bayesian priors.

(v) *Balancing the effects of bottom-up vs top-down*
Focusing the visual inspection on a target object is obviously beneficial as it rejects the erroneous focus points belonging to the background or irrelevant objects. Nevertheless, such an approach can be seen as too "conservative" [50] regarding the information recollection, and the system might neglect interesting data out of its visual exploration area. This is consistent with observations highlighting that global context gives hints to guide the eye gazing of humans patients [3]. Yet another open question is whether the explicit goal of the agent should be to seek an exclusive target, or more broadly put, to expand its knowledge of the visual environment based on global contextual low-level features [86] or an intrinsic motivation [87, 88].

As stated in [89], we do agree that an active approach for perception systems is crucial today for an autonomous agent, likely to face large-scaled cognition tasks in an open and complex world. Understanding the world involves interacting with it. Therefore an active perception is to be understood as more than a data acquisition [90].

Perceiving and recognizing an object is fundamentally a dynamic process where actions (at different levels) are at the forefront. A sensori-motor approach for visual recognition is important to learn correlations between actions and sensations [91, 92, 93].

7. References

[1] Treisman A M and Gelade G 1980 *Cognitive psychology* **12** 97–136

[2] Wolfe J M, Cave K R and Franzel S L 1989 *Journal of Experimental Psychology: Human perception and performance* **15** 419

[3] Torralba A, Oliva A, Castelhano M S and Henderson J M 2006 *Psychological review* **113** 766

[4] Sheinberg D L and Logothetis N K 2001 *J. Neurosci.* **21** 1340–1350 ISSN 0270-6474

[5] Eliasmith C and Anderson C H 2002 *Neural Engineering (Computational Neuroscience Series): Computational, Representation, and Dynamics in Neurobiological Systems* (Cambridge, MA, USA: MIT Press) ISBN 0262050714

[6] Rensink R A 2000 *Visual Cognition* **7** 17–42 ISSN 1350-6285

[7] Yarbus A L 1967 Eye movements during perception of complex objects *Eye movements and vision* (Springer) pp 171–211

[8] Norton D and Stark L 1991 *Scientific American* **224** 34–43

[9] Grossberg S, Mingolla E and Ross W D 1994 *Psychological Review* **101** 470

[10] Ungerleider L G and Haxby J V 1994 *Curr. Opin. Neurobiol.* **4** 157–165 ISSN 0959-4388 URL `https://www.ncbi.nlm.nih.gov/pubmed/8038571`

[11] Katsuki F and Constantinidis C 2014 *The Neuroscientist* **20** 509–521

[12] Sermanet P, Eigen D, Zhang X, Mathieu M, Fergus R and LeCun Y 2014 Overfeat: Integrated recognition, localization and detection using convolutional networks. 2nd international conference on learning representations, iclr 2014 2nd International Conference on Learning Representations, ICLR 2014 ; Conference date: 14-04-2014 Through 16-04-2014

[13] Liu W, Anguelov D, Erhan D, Szegedy C, Reed S, Fu C Y and Berg A C 2016 Ssd: Single shot multibox detector *Computer Vision – ECCV 2016* ed Leibe B, Matas J, Sebe N and Welling M (Cham: Springer International Publishing) pp 21–37 ISBN 978-3-319-46448-0

[14] Redmon J, Divvala S K, Girshick R B and Farhadi A 2016 You only look once: Unified, real-time object detection *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)* pp 779–788

[15] Ren S, He K, Girshick R and Sun J 2015 Faster r-cnn: Towards real-time object detection with region proposal networks *Advances in neural information processing systems* pp 91–99

[16] Szegedy C, Liu W, Jia Y, Sermanet P, Reed S, Anguelov D, Erhan D, Vanhoucke V and Rabinovich A 2015 Going deeper with convolutions *Proceedings of the IEEE conference on computer vision and pattern recognition* pp 1–9

[17] Howard A G, Zhu M, Chen B, Kalenichenko D, Wang W, Weyand T, Andreetto M and Adam H 2017 *CoRR* **abs/1704.04861** (*Preprint* `1704.04861`)

[18] Iandola F N, Moskewicz M W, Ashraf K, Han S, Dally W J and Keutzer K 2016 *CoRR* **abs/1602.07360** (*Preprint* `1602.07360`) URL `http://arxiv.org/abs/1602.07360`

[19] Simonyan K and Zisserman A 2014 *CoRR* **abs/1409.1556**

[20] He K, Zhang X, Ren S and Sun J 2016 Deep residual learning for image recognition *Proceedings of the IEEE conference on computer vision and pattern recognition* pp 770–778

[21] Uijlings J R R, van de Sande K E A, Gevers T and Smeulders A W M 2013 *International Journal of Computer Vision* **104** 154–171 URL `https://ivi.fnwi.uva.nl/isis/publications/2013/UijlingsIJCV2013`

[22] Girshick R, Donahue J, Darrell T and Malik J 2014 Rich feature hierarchies for accurate object detection and semantic segmentation *2014 IEEE Conference on Computer Vision and Pattern Recognition* pp 580–587

[23] Harris C G, Stephens M *et al.* 1988 A combined corner and edge detector. *Alvey vision conference* vol 15 (Citeseer) pp 10–5244
[24] Bay H, Tuytelaars T and Van Gool L 2006 Surf: Speeded up robust features *European conference on computer vision* (Springer) pp 404–417
[25] Rublee E, Rabaud V, Konolige K and Bradski G 2011 Orb: An efficient alternative to sift or surf *Computer Vision (ICCV), 2011 IEEE international conference on* (IEEE) pp 2564–2571
[26] Lowe D G 1999 Object recognition from local scale-invariant features *Computer vision, 1999. The proceedings of the seventh IEEE international conference on* vol 2 (Ieee) pp 1150–1157
[27] Leibe B, Leonardis A and Schiele B 2008 *International Journal of Computer Vision* **77** 259–289 ISSN 1573-1405 URL https://doi.org/10.1007/s11263-007-0095-3
[28] Borji A and Itti L 2013 *IEEE Transactions on Pattern Analysis and Machine Intelligence* **35** 185–207 ISSN 0162-8828
[29] Itti L, Koch C and Niebur E 1998 *IEEE Transactions on Pattern Analysis and Machine Intelligence* **20** 1254–1259 ISSN 0162-8828
[30] Itti L and Koch C 2001 *Nat. Rev. Neurosci.* **2** 194 ISSN 1471-0048
[31] Le Meur O and Liu Z 2015 *Vision research* **116** 152–164
[32] Najemnik J and Geisler W S 2005 *Nature* **434** 387
[33] Friston K, Adams R, Perrinet L and Breakspear M 2012 *Frontiers in psychology* **3** 151
[34] Ablavatski A, Lu S and Cai J 2017 Enriched deep recurrent visual attention model for multiple object recognition *2017 IEEE Winter Conference on Applications of Computer Vision (WACV)* pp 971–978
[35] Mnih V, Heess N, Graves A and kavukcuoglu k 2014 Recurrent models of visual attention *Advances in Neural Information Processing Systems 27* ed Ghahramani Z, Welling M, Cortes C, Lawrence N D and Weinberger K Q (Curran Associates, Inc.) pp 2204–2212 URL http://papers.nips.cc/paper/5542-recurrent-models-of-visual-attention.pdf
[36] Riesenhuber M and Poggio T 2000 *Nat. Neurosci.* **3** 1199 ISSN 1546-1726
[37] Edelman S and Poggio T 1991 *Current opinion in neurobiology* **1** 270–273
[38] Andreopoulos A and Tsotsos J K 2013 *Computer Vision and Image Understanding* **117** 827–891 ISSN 1077-3142
[39] O'Keefe J and Nadel L 1979 *Behavioral and Brain Sciences* **2** 487–494 ISSN 1469-1825
[40] Rolls E T and O'Mara S M 2004 *Hippocampus* **5** 409–424 ISSN 1098-1063
[41] Tsotsos J K 1990 *Analyzing vision at the complexity level* vol 13 (Cambridge University Press)
[42] Tootell R B, Silverman M S, Switkes E and De Valois R L 1982 *Science* **218** 902–904
[43] Weiman C F R and Chaikin G 1979 *Computer Graphics and Image Processing* **11** 197–226 ISSN 0146-664X
[44] Schwartz E L 1980 *Vision research* **20** 645–669
[45] Traver V J and Bernardino A 2010 *Robotics and Autonomous Systems* **58** 378–398
[46] de Figueiredo R P, Bernardino A, Santos-Victor J and Araújo H 2018 *Autonomous Robots* **42** 459–476
[47] Biederman I 1987 *Psychological review* **94** 115
[48] Andreopoulos A and Tsotsos J K 2010 A theory of active object localization *2009 IEEE 12th International Conference on Computer Vision (ICCV)* vol 00 pp 903–910 ISSN 1550-5499 URL doi.ieeecomputersociety.org/10.1109/ICCV.2009.5459332
[49] Gottlieb J 2007 *Neuron* **53** 9 – 16 ISSN 0896-6273
[50] Daucé E 2018 *Frontiers in neurorobotics* **12** 76
[51] Rolls E T, Aggelopoulos N C and Zheng F 2003 *Journal of Neuroscience* **23** 339–348
[52] Posner M I, Rafal R D, Choate L S and Vaughan J 1985 *Cognitive Neuropsychology* **2** 211–228
[53] Klein R M 2000 *Trends in cognitive sciences* **4** 138–147
[54] Tipper S P, Driver J and Weaver B 1991 *The Quarterly Journal of Experimental Psychology Section A* **43** 289–298 (*Preprint* https://doi.org/10.1080/14640749108400971) URL https://doi.org/10.1080/14640749108400971

[55] Goodale M A and Milner A D 1992 *Trends Neurosci.* **15** 20–25 ISSN 0166-2236 URL https://www.ncbi.nlm.nih.gov/pubmed/1374953
[56] Gaussier P and Zrehen S 1995 *Robotics and Autonomous Systems* **16** 291–320
[57] Boucenna S, Cohen D, Meltzoff A N, Gaussier P and Chetouani M 2016 *Scientific reports* **6** 19908
[58] Jauffret A, Cuperlier N and Gaussier P 2015 *Frontiers in neurorobotics* **9** 1
[59] Cuperlier N, Quoy M and Gaussier P 2007 *Frontiers in neurorobotics* **1** 3
[60] Lepretre S, Gaussier P and Cocquerez J P 2000 From navigation to active object recognition *The Sixth Conference on Simulation for Adaptative Behavior* (SAB)
[61] Rothganger F, Lazebnik S, Schmid C and Ponce J 2006 *International Journal of Computer Vision* **66** 231–259 ISSN 1573-1405 URL https://doi.org/10.1007/s11263-005-3674-1
[62] Powers D M 2011 *Journal of Machine Learning Technologies* **2** pp 37–63
[63] Lin T, Goyal P, Girshick R, He K and Dollar P 2018 *IEEE Transactions on Pattern Analysis and Machine Intelligence* 1–1 ISSN 0162-8828
[64] Jia Y, Shelhamer E, Donahue J, Karayev S, Long J, Girshick R, Guadarrama S and Darrell T 2014 Caffe: Convolutional architecture for fast feature embedding *Proceedings of the 22nd ACM international conference on Multimedia* (ACM) pp 675–678
[65] Lin T Y, Maire M, Belongie S, Hays J, Perona P, Ramanan D, Dollár P and Zitnick C L 2014 Microsoft coco: Common objects in context *Computer Vision – ECCV 2014* ed Fleet D, Pajdla T, Schiele B and Tuytelaars T (Cham: Springer International Publishing) pp 740–755 ISBN 978-3-319-10602-1
[66] Erhan D, Szegedy C, Toshev A and Anguelov D 2014 Scalable object detection using deep neural networks *Proceedings of the IEEE conference on computer vision and pattern recognition* pp 2147–2154
[67] Bicanski A and Burgess N 2019 *Current Biology* **29** 979–990
[68] Forssén P E, Meger D, Lai K, Helmer S, Little J J and Lowe D G 2008 Informed visual search: Combining attention and object recognition *Robotics and automation, 2008. icra 2008. ieee international conference on* (IEEE) pp 935–942
[69] Potthast C, Breitenmoser A, Sha F and Sukhatme G S 2016 *Robotics and Autonomous Systems* **84** 31–47
[70] Liu R, Lehman J, Molino P, Such F P, Frank E, Sergeev A and Yosinski J 2018 An intriguing failing of convolutional neural networks and the coordconv solution *Advances in Neural Information Processing Systems* pp 9628–9639
[71] Sabour S, Frosst N and Hinton G E 2017 Dynamic routing between capsules *Advances in Neural Information Processing Systems* pp 3856–3866
[72] Sabour S, Frosst N and Hinton G 2018 Matrix capsules with em routing *6th International Conference on Learning Representations, ICLR*
[73] Viola P and Jones M 2001 Robust real-time object detection *International Journal of Computer Vision*
[74] Bülthoff H H and Edelman S 1992 *Proceedings of the National Academy of Sciences* **89** 60–64
[75] Zhang C, Bengio S, Hardt M, Recht B and Vinyals O 2017 Understanding deep learning requires rethinking generalization *5th International Conference on Learning Representations, ICLR 2017, Toulon, France, April 24-26, 2017, Conference Track Proceedings* URL https://openreview.net/forum?id=Sy8gdB9xx
[76] Arathorn D 2001 *Electronics Letters* **37** 164–166
[77] Arathorn D 2006 A cortically-plausible inverse problem solving method applied to recognizing static and kinematic 3d objects *Advances in neural information processing systems* pp 59–66
[78] Poggio T and Bizzi E 2004 *Nature* **431** 768
[79] Olshausen B A and Field D J 1996 *Nature* **381** 607
[80] Olshausen B A and Field D J 1997 *Vision research* **37** 3311–3325
[81] Zelinsky G J 2008 *Psychological review* **115** 787
[82] de Lange F P, Heilbron M and Kok P 2018 *Trends in Cognitive Sciences* **22** 764 – 779 ISSN 1364-

6613 URL http://www.sciencedirect.com/science/article/pii/S1364661318301396
[83] Rao R P N and Ballard D H 1999 *Nat. Neurosci.* **2** 79 ISSN 1546-1726
[84] Najemnik J and Geisler W S 2009 *Vision research* **49** 1286–1294
[85] Friston K, FitzGerald T, Rigoli F, Schwartenbeck P and Pezzulo G 2017 *Neural Computation* **29** 1–49
[86] Itti L and Baldi P 2009 *Vision research* **49** 1295–1306
[87] Oudeyer P Y and Kaplan F 2008 How can we define intrinsic motivation? *Proceedings of the 8th International Conference on Epigenetic Robotics: Modeling Cognitive Development in Robotic Systems, Lund University Cognitive Studies, Lund: LUCS, Brighton* (Lund University Cognitive Studies, Lund: LUCS, Brighton)
[88] Craye C, Filliat D and Goudou J F 2018 *IEEE Transactions on Cognitive and Developmental Systems*
[89] Bajcsy R, Aloimonos Y and Tsotsos J K 2018 *Autonomous Robots* **42** 177–196
[90] Bajcsy R 1988 *Proceedings of the IEEE* **76** 966–1005
[91] Maillard M, Gapenne O, Hafemeister L and Gaussier P 2005 Perception as a dynamical sensori-motor attraction basin *European Conference on Artificial Life* (Springer) pp 37–46
[92] Gaussier P 2001 Toward a cognitive system algebra: A perception/action perspective *European workshop on learning robots (EWRL)* (Citeseer) pp 88–100
[93] Jauffret A, Cuperlier N, Tarroux P and Gaussier P 2013 *Frontiers in neurorobotics* **7** 16