Active Perception: Scene Exploration using Foveal Vision

Luís Doutor Simões
luis.d.simoes@tecnico.ulisboa.pt

Instituto Superior Técnico, Lisboa, Portugal

September 2021

Abstract

Active perception and foveal vision are the foundations of our visual system. While foveal vision reduces the amount of information to process at any time instance, active perception will direct the eyes to promising parts of the visual field. Together, they allow a detailed perception of the objects on the environment with limited neuronal processing resources. We develop a method that combines both concepts to explore and identify all the objects on an image with the least number of gaze shifts. A foveal sensor will scan the image sequentially and create a semantic map of the scene, choosing at each step the location with higher information gain, regarding the identification of the objects. Our framework uses the foveated images as input to a state-of-the-art object detector, whose scores are modelled by a Dirichlet distribution that depends on the distance to the fovea, denoted Foveal Observation Model. After each new saccade, this Model is used to perform a Sequential Fusion of the detection scores in a global map. With the updated distributions at each map point, a decision based on information theoretic measures is made to find the next-best-viewpoint that maximizes our knowledge of the world. Despite the blur, we show that it is possible to combine foveated images with state-of-the-art object detectors using our proposed models. Furthermore, our models not only improve the identification of objects by 2-3%, but also reduce 3x (in average) the number of required gaze shifts to achieve similar performances against randomly choosing the next viewpoint.
Keywords: Active Perception; Foveal Vision; Object Detection; Active Object Search; Fusion of Classifiers

1. Introduction

Central vision (or foveal vision) is an indispensable feature of the human eye allowing to perform activities which require high-resolution visual details, in contrast with peripheral vision where the resolution is much lower (blurred image).

So, why are our eyes divided in these two regions? It would be reasonable to think that having a wider central vision could greatly improve our survival. But in fact, human eyes are built the other way around. The fovea comprises less than 1% of retinal size, but takes up over 50% of the cortex [10], thus one can imagine that our brain would have to be impractically large to handle the full visual field at high resolution.

However, since the amount of information is greatly reduced by the foveation mechanism, one could think that, to analyse a scene, it would just require a gaze shift to every location, and extract the information obtained by the fovea. Still, scanning the entire scene would require an unbearable amount of time. Nevertheless, the fovea does not need to cover the entire scene, the peripheral vision also extracts some useful information to guide the eyes to visit unexplored places where there is a high probability of existing objects, given all the acquired information.

Just like human vision, many computer vision applications are constrained by the involved computational effort, specially when implemented on artificial intelligent agents whose tasks depend on the analysis, in real-time, of their surroundings. Hence, urges the need to develop models capable of filtering and fusing information, ignoring what is not relevant for the task in hands. This is where the Active Perception models, combined with foveal vision, come to the picture. Active perception selectively chooses new targets for the acquisition of information based on the knowledge that the agent has about the current state of the world and what is promising or not to complete a certain task.

Although there have been a large amount of research and developments on attention and visual search models (as in [4], [3], [1] and [11]), there is still a long way to go, specially regarding the modeling of the mechanisms that help the decision of where to shift the gaze to. Besides, at the extent of our knowledge, there are no attempts on com-

bining state-of-the-art object detection mechanisms and active perception methods to perform a scene exploration task using foveal vision.

1.1. Problem Definition

The main goal of this project is to implement a model to optimize the exploration of a scene, gathering as much information as possible about all the objects, in the least amount of gaze shifts, using foveal vision.

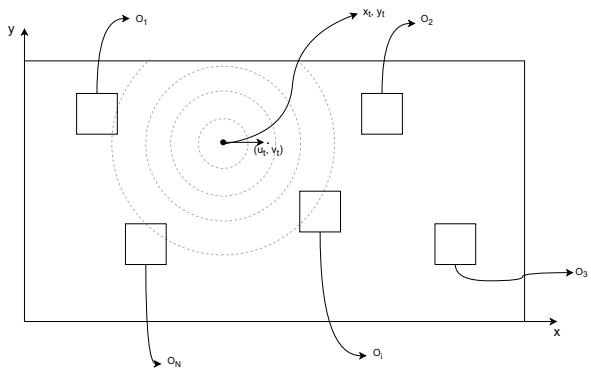


Figure 1: Scene/image representation where the squares represent the objects O and the center of the dashed circles represents the focal point (x_t, y_t).

So, let's start by considering $\mathcal{C}$ as the set of possible classes of objects O_m

$$\mathcal{C} = \{c_o, c_1, ..., c_K\} \tag{1}$$

where K is the number of classes, and c_0 is the label of the background class.

The set of detected objects $\mathcal{I}_t$ seen by a detector algorithm (may be different from the actual number of objects), at instant t is given by

$$\mathcal{I}_t = \{I_{t,l}\}, \quad l = 1, ..., L_t \tag{2}$$

where

$$I_{t,l} = (\boldsymbol{B}_{t,l}, \boldsymbol{S}_{t,l}) \tag{3}$$

being $\boldsymbol{B}_{t,l}$ a bounding box, which is an array containing the location and size of the object, and $\boldsymbol{S}_{t,l}$ an array of confidence scores. The position of the bounding box $\boldsymbol{B}_{t,l}$ can be given by the local coordinates $(u_{t,l}, v_{t,l})$ representing the relative position of the center of the bounding box to the focal point (x_t, y_t). On the other hand, the confidence scores $\boldsymbol{S}_{t,l}$ contain the probability of a given detection $I_{t,l}$ belonging to each of the K classes of objects $\mathcal{C}$ for which the detector was trained to detect:

$$\boldsymbol{S}_{t,l} = [s_{t,l,1}, s_{t,l,2}, ..., s_{t,l,K}]^T, \quad 0 \leq s_{t,l,j} \leq 1 \tag{4}$$

The confidence scores $\boldsymbol{S}_{t,l}$ are in the probability simplex after normalizing the probabilities to sum to one, $\sum_{k=1}^{K} s_{t,l,k} = 1$ and $s_{t,l,k} \geq 0$ for all $k \in \{1, ..., K\}$.

After having the output of the detections on the resulting foveated images of each saccade, we first need to build an Observation Model that models how the detections and their confidence scores vary depending on their relative position to the retina. The Observation Model is then defined for each detection $I_{t,l}$ as the distribution of its confidence scores $\boldsymbol{S}_{t,l}$, given the distance to the fovea $d_{t,l} = \|(u_{t,l}, v_{t,l})\|$, for each possible object class label c_k, :

$$p(\boldsymbol{S}_{t,l}|c_k, d_{t,l}) \tag{5}$$

Secondly, a world map has to accumulate over time the knowledge that the observations provide, for our application we can consider a body centered 2D map of the surroundings. Therefore, a Fusion Model is required to update the map with new observations, after each saccade

$$\boldsymbol{M}_t(x, y) = P(C_{x,y}|I_{0:t,f_{t,l}(x,y)}, x_{0:t}, y_{0:t}) \tag{6}$$

where $\boldsymbol{M}_t$ can be seen as the map information (state) at iteration t, containing at each pixel (x, y) a vector of parameters that encode the probability distribution of the fusion of all observations that overlap the pixel (x, y), which are given by the function $f_{t,l}(x, y)$, where $C_{x,y} \in \mathcal{C}$ is the class label that we want to estimate

Having now the updated map information, in order to make a full exploration in the least number of gazes, an Active Perception method that chooses the point to look next that maximizes the gain of information about the scene has to be implemented

$$x^*, y^* = \underset{x,y}{\text{argmax}}\ F(x, y, \boldsymbol{M}) \tag{7}$$

where $F(x, y, \boldsymbol{M})$ corresponds to the gain of information or the loss of confusion.

2. Background & Related Work

In this section we will review related work on active perception and integration with foveal vision, as well as recent work on how to fuse observations in order to keep updating the knowledge that the agent has of the world.

2.1. Active Perception

Machine Learning techniques often rely on huge amounts of labeled data. The data is then processed by a training algorithm, which optimizes the parameters to perform the task for what it was designed. One constrain of these machine learning techniques, and perhaps the biggest one, is the insufficient amount of available data to train the algorithm, and the time it would take to process it.

To overcome this issue, active learning began to emerge as a hot scientific topic. Active Learning is

built upon the principle that the learning algorithm has the ability to choose the data from which it learns, and, this way, if the the data is well chosen, the algorithm can perform better with less training [15].

Active perception is a particular subset of active learning. The agent acquires information directly from the sensors, which is combined with prior knowledge of the world and the current state, to then select the next information to gather [5]. Active perception can be performed with different kinds of sensors and stimulus. The focus of this work is on solving a search problem using visual sensory information, and, therefore, the active perception specialization that will be studied here is given the name of active vision [2]. This problem can be seen as a planning problem, denominated "next best view point".

The choice of the next best view point is made through the use of acquisition functions, where the objective is to choose the point that maximizes a function related to our objective. Figueiredo [7] on his work used three common acquisition functions, the Upper Confidence Bound, the Probability of Improvement and the Expected Improvement. The application of the acquisition functions depends on the type of information acquired by the sensors and others can be derived from these reference ones. On section 3.4 the acquisition functions developed to our specific case will be explained in detail.

Earlier work on actively searching for objects was proposed by Ayedemir [4], with an Active Visual Search strategy considering topological relations between objects. The approach had a major drawback, the amount of prior information needed, which the user had to input whenever a new search was to be performed. Anyway, Aydemir latest work [3], where he added the uncertain semantic of the environment, already provided promising results when compared against humans on performing an object search task on unknown map.

On another approach, a new mechanism combining stereo vision and active perception was purposed by Grotz [8], where a more task-related gaze selection was explored, based on multiple saliency maps. His objective was to reduce the uncertainty associated to the desired object pose to then be able to grab it more efficiently. The detection of the object was made based on the extraction of local features and the uncertainty associated to the object pose was modelled as a Gaussian distribution and updated using a Kalman filter. Nevertheless, the use of local feature detectors greatly reduces the possible complexity of the objects present on the scene.

We were actually inspired by a more recent work developed by Figueiredo [7] where depth information was combined with the uncertainty in stereo matching to perform an active gaze selection method. The objective was to extract the maximum amount of information of the closest object to the camera while updating the world map using foveal mechanisms. Figueiredo's results showed that, with the right parameters, foveal vision would outperform Cartesian, regarding the amount of information extracted. Nevertheless, besides the promising results, the optimization criteria was to choose the closest object, with disregard for the type of object itself.

Following Figueiredo's work [7], an iterative approach combining saliency maps (inspired on Grotz approach [8]) with active perception to improve the detection of objects was proposed by Almeida [1]. Almeida proposed a biological inspired object classification and localization framework combining DCNN with foveal vision. First, a DCNN operates over the foveated image to predict the class labels. Then, a color-based saliency map is used to obtain the object location proposal. At the next iteration, the center of the location proposal is used as the new foveation point, and the process is repeated, in order to try to improve the classification and localization of the object. As in Grotz work, the use of this kind of saliency maps reduces the quality of the localization of the objects as the image gets more complex. Besides that, Almeida's framework considered images with just one object.

Other biological inspired work was performed by Melicio [11], where attention mechanisms were combined with foveal vision to perform image classification. Melicio dropped the model based saliency maps by using a CNN to both detect salient regions and classify the foveated image in just one step. The salient regions outputted by the CNN were then used to shift the foveation point to locations that would potentially improve the classification. Melicio showed that after the gaze shift, the performance improves. In her work the localization of the object is not required, since, as in Almeida's work, the objective was to classify an image containing one central object. Thus, the uncertainty in the detection imposed by the foveal vision did not need to be modelled.

2.2. Approaches on Fusing Observations

The fusion problem consists in, given a set of classification scores for a single pattern (which can be from distinct classifiers that produce observations at the same time, to a single classifier that obtains consecutive measurements for the same pattern but in different time instants), how can one calculate a single global classification score $\boldsymbol{p}$ and/or estimate its distribution. For our specific case the objective here would be to update the world map $\boldsymbol{M}$ enunciated in eq.(6)

As an example, Montesano [13] developed an algorithm that learns local visual descriptors of good grasping points based on a set of trials performed by the robot. The parameters of the corresponding distribution (in this case Beta distribution) are updated as a simple function of the number of successes and failures.

Although we will not have "successes" and "failures" to use Montesano's approach, we can use Figueiredo's [7] sequential Bayesian filtering. Bayesian filtering allows one to accumulate sensor inputs and update the likelihood of a map point being the desired object, at each time instant, assuming that we know the probability distribution of the data.

Considering an arbitrary pixel and that we only have one score vector per instant of time for that pixel $\boldsymbol{S_t}$ (which is not the case, but simplifies the notation), the Bayesian filtering, or Naïve Bayes approach, updates the map distribution $\boldsymbol{M_t}$ by

$$\begin{aligned}\boldsymbol{M}_T = P(C|\boldsymbol{S}_1, ..., \boldsymbol{S}_T) = \frac{1}{Z} P(C) \prod_{t=1}^{T} P(\boldsymbol{S}_t|C) = \\ = \frac{1}{A} P(\boldsymbol{S}_T|C) \boldsymbol{M}_{T-1}\end{aligned} \tag{8}$$

and, individualizing for each class of objects c_j

$$p(c_j|\boldsymbol{S}_1, ..., \boldsymbol{S}_T) = \frac{1}{Z} p(c_j) \prod_{t=1}^{T} s_{t,j} \tag{9}$$

where A and Z are normalizing constants that do not depend on the object class, and $p(c_j)$ is the prior class probability for c_j.

An interesting work on testing the performances of the Naïve Bayes against other fusion methods was performed by Kaplan [9], where a new approach on fusing classifiers was proposed. The fusion method proposed by Kaplan maps the classifications into a Dirichlet distribution with parameters $\boldsymbol{\beta}_t^{x,y}$, being able to take into consideration the uncertainty associated to each classifier when fusing. This is interesting since it allows to store the information of how many updates were done at each map cell. Let's again consider an arbitrary pixel (x, y) and, for simplicity, omit the coordinates from the equations of this section.

First, Kaplan proposes a what he calls naïve approach that instead of adding the number of occurrences to the Dirichlet parameters (as in Montesano approach [13]), the Dirichlet parameters are updated by adding the actual confidence scores

$$\beta_{t+1,k} = \beta_{t,k} + s_{t,k} \tag{10}$$

This approach is actually equivalent to a classical approach named sum rule, considering uniform priors. The sum rule tries to approximate the posterior class probabilities, but instead of multiplying the observations, it sums them, presenting a more robust solution to outliers than the Naïve Bayes, since values close to zero would automatically lead the result of the Naïve Bayes to low probability values. Nevertheless, Kaplan states that this sum rule does not yield a posterior Dirichlet distribution that fits well the actual posterior distribution of $\boldsymbol{p}$.

Kaplan then proposes a new updated, by approximating a Dirichlet distribution to the actual posterior $f(\boldsymbol{p}|\boldsymbol{\beta}, \boldsymbol{S})$ through a moment matching approach. On this new approach on fusing classifiers (Kaplan's update), the Dirichlet parameters are updated by the following equation

$$\beta_{t+1,k} = \frac{\beta_{t,k} \left(1 + \frac{s_{t,k}}{\sum_{j=1}^{K} \beta_{t,j} s_{t,j}}\right)}{1 + \frac{\min_j s_{t,j}}{\sum_{j=1}^{K} \beta_{t,j} s_{t,j}}} \tag{11}$$

It will be interesting to test how this different updates work, as we will be dealing with the uncertainty imposed by the foveal sensor that is not constant: depends on the object location. We will therefore expand Kaplan's work on comparing different fusion methods, to our specific problem.

3. Approach

The proposed project involves the integration of several components (see figure 2). These components will be described and explored throughout this section.

3.1. Foveal Conversion

The Foveal Conversion will collect the image that we wish to explore and use Almeida [1] and Melício [11] model to foveate the collected Cartesian image with the center of the fovea being the one returned by the Gaze Selection block at each iteration. Where they first build a Gaussian scale-space where each level corresponds to a low-passed version of the previous level. Each level has an increasing level of blur, but similar resolution. Then, a Laplacian scale-space is built where the difference between adjacent Gaussian levels is computed, resulting in a set of error images. Finally, each level is multiplied by exponential kernels to emulate a smooth fovea. Their approach creates an image which has a higher resolution around the foveation point, decreasing gradually over the periphery but does not change the pixel size and distribution along the foveated image. Consequently, although simulating foveal vision, this approach does not take advantage of the decrease of resolution over the periphery to reduce computational costs. Anyway, it is a convenient process to analyse the consequences of foveal images in artificial vision and machine learning methods, since it creates an image ready to be processed by a pre-trained CNN for Cartesian images; One just has to resize the image to fit

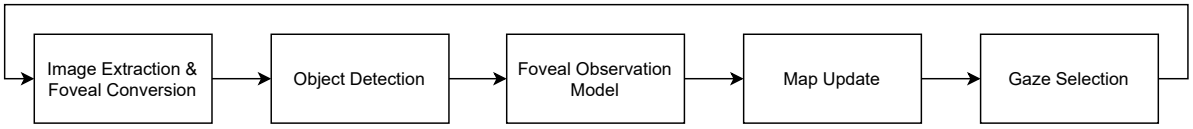


Figure 2: Graphical model of the framework.

the input requirements of the network.

Nevertheless, in order to take full advantage of the possible memory reduction when using foveal vision, a different approach was recently proposed by Siebert & Ozimek [16]. They use a self-similar neural network to define retina sampling locations as described by Clippingdale & Wilson [6]. This approach was only tested for classification tasks and not detection ones, and it would require a re-train of already built detection algorithms. So, since we are more concerned on how to actively explore a scene and search for objects using foveated images, we will leave this interesting approach for future work.

3.2. Object Detection & Foveal Observation Model

The foveated image serves as input to the object detection method (we will use a YOLOv3 [14]), which outputs are modeled by the foveal observation model. The whole process is represented graphically in figure 3.

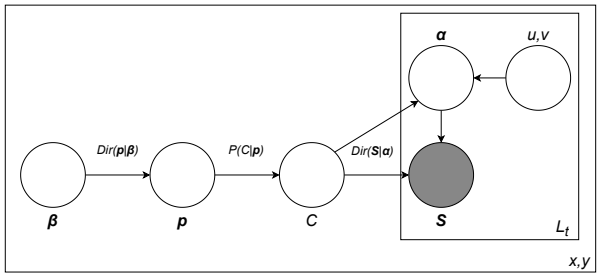


Figure 3: Object detection and Foveal Observation Model diagram.

For each image location, given by the global coordinates (x, y), there is a probability of appearing a given object, represented by a probability vector

$$\boldsymbol{p} = [p_1, p_2, ..., p_K]^T \tag{12}$$

sampled from a Dirichlet prior with parameters $\boldsymbol{\beta}$ that depend on the environment (on our case we are assuming a uniform $\boldsymbol{\beta}$ generates a uniform $\boldsymbol{p}$, i.e. there is no preference for any class of objects). K is the number of possible object classes.

Given $\boldsymbol{p}$, an object represented by the random variable C is sampled, which is then associated to a bounding box.

Given C and the position on the foveated image, our YOLO detector generates, for each instant t and each object detected l $(l = 1, ..., L_t)$, a multinomial score vector $\boldsymbol{S}_{t,l}$. The score vector contains the confidence scores of the detection algorithm for each class of object, as enunciated on eq.(4).

It's important to note that object detection algorithms were built upon the assumption that the input image is Cartesian, meaning that they were not trained to detect the blur imposed by the foveal sensor on the objects, and, therefore, do not know the location of the focal point. That information could be useful to better classify an object affected by the blur on the peripheries, since it would be already expected that the uncertainty and the confusion between object classes would be higher there.

The multinomial score vector $\boldsymbol{S}_{t,l}$ has then a Dirichlet prior that depends on the location of the image (which is expected to have less entropy near the center of the fovea, and higher entropy on the peripheries).

The parameters of the Dirichlet prior that characterizes the uncertainty on the output of the score vector $\boldsymbol{S}_{t,l}$ can be written as:

$$\boldsymbol{\alpha}_{k,d_{t,l}} = [\alpha_{k,d_{t,l},1}, \alpha_{k,d_{t,l},2}, ..., \alpha_{k,d_{t,l},K}]^T \tag{13}$$

that depends on the distance $d_{t,l}$ between the outputted bounding box, and the center of the fovea (x_t, y_t), i.e., depends on the detection local coordinates $(u_{t,l}, v_{t,l})$. These parameters have to be learned from a supervised training set, with the outputs of the classifier obtained on different observation conditions for each of the classes. Each Dirichlet distribution of the Foveal Observation Model was trained using an iterative approach proposed by Minka [12].

The Foveal Observation Model is then composed by a set of different Dirichlet distributions, one for each pair class k and distance level $d_{k,l}$ (on our case 7 different distance levels were considered). Thus, whenever a detection $I_{t,l}$ appears, depending on the distance to the focal point, a Dirichlet distribution is chosen for each class of object $k = 1, ..., K$

$$\begin{aligned}\boldsymbol{S}'_{t,l} =& [s'_{t,l,1}, ..., s'_{t,l,K}] = \\ &= \frac{1}{D}[Dir(\boldsymbol{S}_{t,l}|\boldsymbol{\alpha}_{1,d_{t,l}}), ..., Dir(\boldsymbol{S}_{t,l}|\boldsymbol{\alpha}_{K,d_{t,l}})]^T\end{aligned} \tag{14}$$

Where D is a normalization factor given by $D = \sum_{k=1}^{K} Dir(\boldsymbol{S}_{t,l}|\boldsymbol{\alpha}_{k,d_{t,l}})$, so that $\sum_{k=1}^{K} s'_{t,l,k} = 1$. These new score vector $\boldsymbol{S}'_{t,l}$ is expected to have less confusion than the ones outputted by the detector.

3.3. Fusion Model

In order to simplify the exploration, our world corresponds to a single image, where we wish to correctly detect and classify every object on the least number of gaze shifts. Therefore, the information obtained every time the algorithm moves the eyes has to be stored in a map and fused with the information already obtained on previous iterations.

As proposed by Kaplan [9] (refer back to section 2.2), the fusion results of both the sum rule and "Kaplan's approach" can be mapped onto a Dirichlet distribution with parameters $\boldsymbol{\beta}$. Therefore, storing on the map these parameters $\boldsymbol{\beta}_t^{x_m,y_m}$, for the current instant t, for each map cell (x_m, y_m) will allow one to not only extract the expected probability of each class of objects k (for $k = 1, ..., K$) on that cell

$$p_{t,k}^{x_m,y_m} = \frac{\beta_{t,k}^{x_m,y_m}}{\sum_{j=1}^{K} \beta_{t,j}^{x_m,y_m}} \tag{15}$$

but also to have more information about the uncertainty of these expected values, since the parameters $\boldsymbol{\beta}$ of the Dirichlet distribution contain more information than the categorical distribution $\boldsymbol{p}$ alone.

Updating the map information with the sum rule can then be done through eq.(10), and, using Kaplan's approach can be done through eq.(11). Following Kaplan's [9] work we defined the initial parameters for each fusion algorithm and map cell as $\beta_k^0 = 0.5$ for $k = 1, .., K$.

The other fusion approach that will be tested on our framework, the Naïve Bayes, can not be modelled by a Dirichlet and therefore each map cell will store the categorical distribution $\boldsymbol{p}$ outputted with the Naïve Bayes approach, instead of the $\boldsymbol{\beta}$ parameters as in the other approaches.

For our specific case, the confidence scores outputted by the YOLO algorithm $\boldsymbol{S}_{t,l}$ would be used on the sum rule equation (eq.(10)) and Kaplan's equation (eq.(11)). Nevertheless, for each map cell (x_m, y_m) there might be 0, 1 or more detections at a given instant of time t, meaning that for that instant of time each fusion process for the map cell (x_m, y_m) is repeated for every detection belonging to the set $\mathcal{I}_t^{x_m,y_m}$, where $\mathcal{I}_t^{x_m,y_m}$ is the set of detections at instant t which bounding boxes intersect with the map cell (x_m, y_m).

Also, a "background" class was appended to each score vector, just as it was a confidence score outputted by the object detector (YOLO only assigns confidence scores to objects). The confidence score of the background was chosen to be the value that the detector would output on every class in the highest uncertainty case, the uniform distribution case ($s_{t,l,K+1} = \frac{1}{K+1}$).

Nevertheless, using Kaplan fusion method (eq.(11)) with simply the output of the detector algorithm would ignore the knowledge that we have about the location of the objects in relation to the center of the fovea. Thus, in order to analyse if this knowledge can improve the performance of a scene exploration, the full observation model has to be considered on a modified version of this fusion method, substituting $\boldsymbol{S}_i^{x_m,y_m}$ for $\boldsymbol{S}_i'^{x_m,y_m}$, following eq.(14).

The scores outputted by the foveal observation model (eq.(14)) will serve as input to the Naïve Bayes update (eq.(9)) and to a modified version of Kaplan's approach (eq.(11)), which we will call the "modified Kaplan approach", and will be compared to the classical sum rule (eq.(10)) and Kaplan approach (eq.(11)).

3.4. Active Perception - Gaze Selection

We want to predict what is the next focal point that minimizes the confusion on the map. In this work three common metric to measure the map confusion will be tested: the KL divergence, the entropy and the difference between the two classes with highest probability (difference between two peaks).

The KL divergence is computed over the parameters of the Dirichlet distributions that characterizes the map cells ($\boldsymbol{\beta}$), and measures how different they are from its initial state. Whilst the entropy and the difference between two peaks are computed over the expected values ($\boldsymbol{p}$) of those distributions.

Having metrics that can measure the amount of uncertainty/confusion on each map cell, the best predicted next view point is computed by *acquisition functions* that have to predict the global (average) uncertainty of the map if the focal point changed to another pixel of the image. The *acquisition functions* considered in this work aim to find the cell that minimizes the average map uncertainty with each of the the metrics above.

For the KL Divergence, one aims to maximize the average KL Divergence of the map:

$$(x_m^*, y_m^*) = \underset{i,j}{\operatorname{argmax}} \sum_{x_m=1}^{X} \sum_{y_m=1}^{Y} E^{ij} \left\{ D_{KL}^{x_m,y_m,(t+1)} \right\} \tag{16}$$

where $E^{ij}\{.\}$ corresponds to the expected value for a fovea centered on (i, j).

Using the Classification Entropy metric, one wishes to minimize the average entropy of the map. Thus, following the same notation as above:

$$(x_m^*, y_m^*) = \underset{i,j}{\operatorname{argmax}} \left\{ - \sum_{x_m=1}^{X} \sum_{y_m=1}^{Y} E^{ij} \left\{ Entr^{x_m,y_m,(t+1)} \right\} \right\} \tag{17}$$

For the Difference between Two Peaks, one wishes to maximize the absolute gain of this metric:

$$(x_m^*, y_m^*) = \underset{i,j}{\text{argmax}} \underset{x_m,y_m}{max} \left| D_{2Peaks}^{x_m,y_m,t} - E^{ij}\left\{ D_{2Peaks}^{x_m,y_m,(t+1)} \right\} \right| \quad (18)$$

Predicting the resulting detections and updates of the map, if the algorithm shifts the gaze to a certain position, is possible due to knowing that the fovea will have a higher resolution than the peripheries. Taking advantage of the distance to the center of the fovea to try to predict which objects are where is exactly what the Foveal Observation Model does. Thus, by modeling the current expected values of the state of each map cell, one can predict the evolution of the map for each hypothetical focal point. This is something that can not be done without the Foveal Observation Model.

4. Results

For this section, experiments to validate and analyse the performance of each of the models used on this work were performed.

4.1. Foveal Observation Model Validity

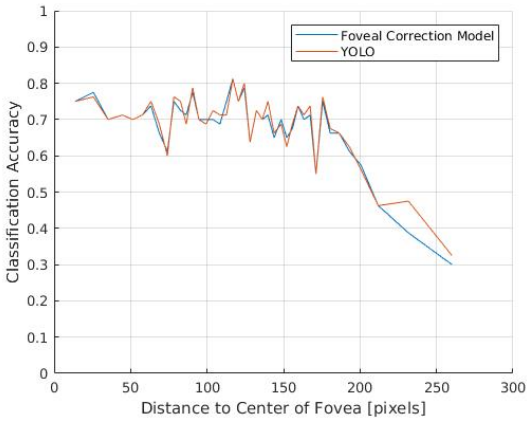
The foveal observation step is supposed to take advantage of the relative position of the objects to the center of the fovea, in order to model the uncertainty imposed by the blur on the peripheries on the output scores of the detected objects.

For this comparison, several random images were taken (from the COCO dataset) and foveated using randomly chosen focal points. The correspondence between the classification outputs and the ground-truth objects was done by finding bounding boxes with an IoU greater than 30% with the ground-truth information.

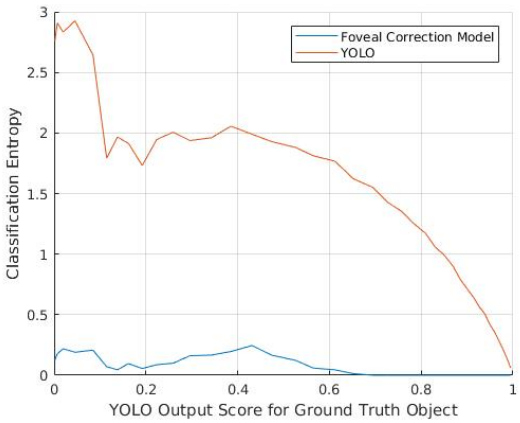
On figure 4 one has the comparison of the classification performance with and without modelling the confidence scores with the foveal observation model. The metrics are an average over every image.

The accuracy lines (figure 4(a)) are very similar, meaning that modelling the detections with the observation model does not impose a drop on performance, proving the validity of the foveal observation model. Moreover, it reduces the uncertainty that the algorithm has on the detection, as one can see on figure 4(b).

Figure 4(b) shows that low confidence scores directly outputted by the detector (*p.e.* detections affected by the blur on the peripheries) have a high degree of entropy , but when these scores are modelled by the observation model, the entropy of the confidence score vector is much lower. The model tries to find the object, for that distance, that better fits the distribution of the scores, even if there is a big confusion among some of the classes, to present a more certain classification. Although this does not prove to be an improvement on a 1-step classification approach (as seen on figure 4(a)), it will still be useful for the multi-step classification approach that we are taking.



(a) Accuracy comparison.



(b) Entropy comparison.

Figure 4: Performance comparison between using (blue) and not using (red) the observation model

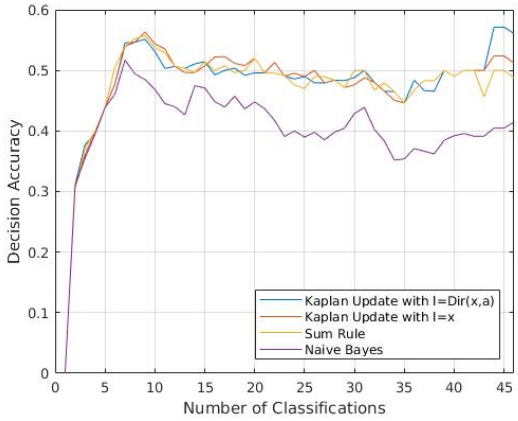
4.2. Fusion Model Performance

Having analysed the performance and validity of the foveal observation model, it's now time to check the performance when using this model to update the information on the map, at each iteration.

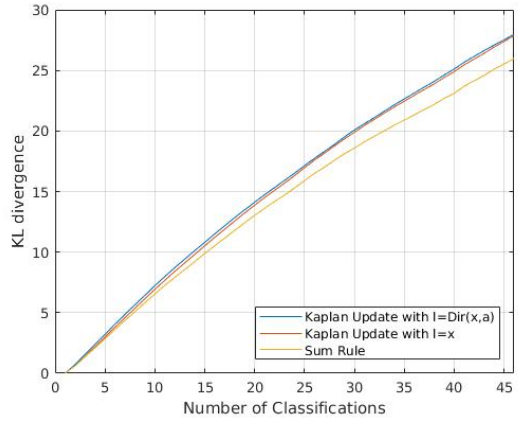
Four different fusion methods were implemented to update, at each iteration, the current state of each map cell - Naïve Bayes (eq.(9)), Kaplan Update (eq.(11)), Modified Kaplan Update (eq.(11) with eq.(14)), and Sum Rule (eq.(10)). On this section, the performance of each method will be tested when fusing detections on foveated images.

The values used in the plots to evaluate this experiment correspond to the average of the accuracy and uncertainty metrics over a set of 50 images from the COCO database, where each one was foveated sequentially with random focal points. The random

exploration approach will be used instead of an active one, allowing to isolate the performance of the fusion methods without considering the gaze selection step. The ground-truth object for each classification is again chosen with the 30% IoU threshold.



(a) Average Accuracy evolution.



(b) Average KL divergence evolution.

Figure 5: Time-wise analysis of the fusion algorithms as new bounding boxes are detected.

On figure 5(a) one can see how the average accuracy evolves as new bounding boxes are detected (each bounding box corresponds to one classification). It is possible to note that every algorithm achieves a similar performance on the accuracy, except for the Naïve Bayes, where the performance is lower. Also, due to the drastic entropy reduction imposed by the foveal observation model, the Naïve Bayes approach was greatly affected by most of the classes having a score closer to zero, outputting no uncertainty upon fusing different observation. That is why the Naïve Bayes was not considered for further analysis (since we need the uncertainty to predict the next view point) and it is not represented on figure 5(b).

As for the other algorithms, both the expected value and the uncertainty have a positive evolution as new bounding boxes are fused, where the Kaplan updates present slightly better results, as in Kaplan experiments [9].

4.3. Active Gaze Selection Performance

Knowing what is the most promising next view point is the key to achieve better performances than, per example, choosing a random point at each iteration.

Let's start by comparing the performance of each acquisition function used in this work (section 2.1): Absolute Gain on the Difference between Two Peaks (eq.(18)), KL Divergence Gain (eq.(16)), and Classification Entropy Loss (eq.(17)).

For this experiment, 150 random images from the COCO dataset were used and each one foveated 10 times. The foveation points were chosen by the acquisition functions and differ from method to method, nevertheless, the starting focal point was randomly chosen for each different image, but is the same on every method. The detections at each iteration were used to update the information of the map using the Modified Kaplan Update (eq.(11) with eq.(14)), since only the Modified Kaplan approach depends on the distance of the detection to the center of the fovea, and, therefore, it is the only one that can predict the evolution of the map in order to choose the most promising view point.

On figure 6 one can see a clear difference between using the Classification Entropy Loss and the other acquisition functions. Although both the entropy and the KL Divergence measure similarly the amount of confusion on a map cell, the KL Divergence combines that confusion with the amount of updates done in that particular cell, more updates mean less uncertainty even if the probability of every class is the same. We can then say that the KL Divergence is the most suited metric to measure the uncertainty of the Dirichlet distributions that characterize the state of the map.

It is now time to compare the benefits of active gaze selection with respect to random search. The experiment is the same as before but now the Modified Kaplan fusion method is tested against all the other fusion methods, which choose the next focal point randomly.

The results of figure 7 are promising. One can immediately notice that the Modified Kaplan update jointly with the best Active Perception method analysed achieves better F1-Score at almost every iteration than all other fusion algorithms when choosing randomly the next focal point.

One other important aspect is the growth rate of the performance on classifying the objects on the image. Since the goal is to find and classify every object on the image, in the least number of gaze

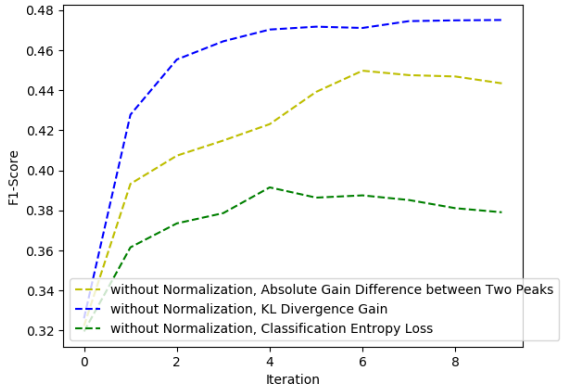


Figure 6: Comparison between the F1-Scores using the three different acquisition functions.

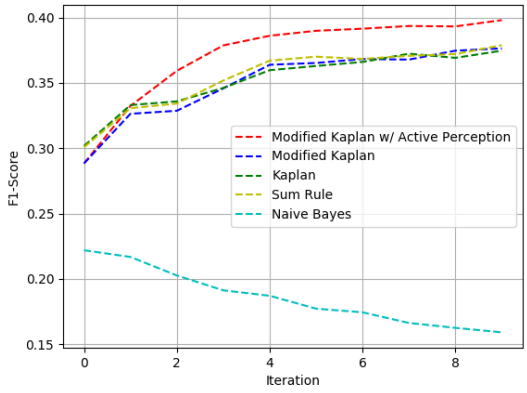


Figure 7: F1-Score comparison of the Modified Kaplan using the acquisition function "KL Divergence Gain" (red), against the ones choosing the focal point randomly.

shifts, analysing how fast the algorithm can detect and correctly classify most of the objects is a key factor. As one can see, choosing the next focal point by maximizing the predicted gain on the average KL divergence of the map, achieves an F1-Score around the third iteration that can not be surpassed by any of the methods that use random search.

Besides the improved growth rate, we can also see on figure 7 that choosing the next focal point by maximizing the KL Divergence Gain contributes to an overall performance improvement (on average) of around 2-3% on the F1-Scores after the 10 iterations of the experiment.

5. Conclusions

In this work we propose a computational framework, inspired by human vision, that incorporates the combination of foveal vision and a state-of-the-art object detector with recent approaches on fusion of classifiers, to perform an active exploration for objects.

The main goal was to find and correctly classify as many objects as possible in one image, in the least number of gaze shifts. For this purpose, the work was divided in three major components. The Foveal Observation Model (section 3.2), the fusion of incoming observations (section 3.3), and the prediction of the next best view point (section 3.4).

Regarding the first component, object detectors were built to locate and classify objects on Cartesian images, thus, one of our contributions was to train, develop, and analyse a Foveal Observation Model that post-processes the results of the object detector, taking advantage of the confusion imposed by the blur on the periphery of the image to try to classify the objects in one passage.

The classification performance of the Foveal Observation Model was validated in our tests (as presented on section 4.1), reducing the uncertainty imposed on the classification scores while achieving a similar accuracy when compared to the object detector itself. From these results, we can conclude that the confusion between classes, due to the increasing blur as we go to the peripheries of the fovea, can be modeled. Moreover, the Observation Model could make good predictions about the evolution of the map on the next iteration (section 4.3), depending on the location of the fovea, based only on the current knowledge. This is one of the most important features and contributions of this Observation Model, since it allows one to use this predictions to then choose the most promising point where to look next.

About updating the map information, we proposed a modified version of Kaplan's fusion algorithm, combining it with the outputs of the Foveal Observation Model instead of using directly the outputs of the object detector. We concluded on section 4.2 that the outputs of the Foveal Observation Model remain valid when combined with the fusion algorithm, and that the greediness of the classification using the observation model is compensated by the limits imposed by the fusion algorithm on each update, something that doesn't happen when fusing these outputs with the Naïve Bayes approach.

The results obtained on the last component, the active gaze selection, are significant, and validate the proposed framework as being the first exploration algorithm with foveal vision that takes advantage of the performance of a state-of-the-art detector. The algorithm achieved a performance more than three times faster by trying to shift the gaze to the location that maximizes the KL divergence gain, and also contribute with an overall improvement of 2-3% of the performance (F1-Score), than when choosing randomly the next focal point (sec-

tion 4.3). The results show that it is possible to take advantage of the uncertainty imposed by this kind of images to optimize the exploration of a scene.

We can finally conclude that this work contributes with a promising new approach on active exploration, since it is a first step on taking advantage of the performance of a state-of-the-art object detector, trained on Cartesian images, to develop a searching algorithm using foveal vision. By modelling the uncertainty imposed by the image on the detections we showed that it was possible to perform a search for objects on a given environment without resorting to more specific and limited heuristics.

For future work it would be interesting to expand the search to a real-world scenario, instead of being restricted to an image, and also analyse if the reduction of the required number of saccades can be translated in computational gains, when compared with a search done using full-resolution vision.

References

[1] A. F. Almeida, R. Figueiredo, A. Bernardino, and J. Santos-Victor. Deep Networks for Human Visual Attention: A Hybrid Model Using Foveal Vision. In A. Ollero, A. Sanfeliu, L. Montano, N. Lau, and C. Cardeira, editors, *ROBOT 2017: Third Iberian Robotics Conference*, pages 117–128, Cham, 2018. Springer International Publishing.

[2] J. Aloimonos, I. Weiss, and A. Bandyopadhyay. Active vision. *International Journal of Computer Vision*, 1(4):333–356, 1988.

[3] A. Aydemir, A. Pronobis, M. Gobelbecker, and P. Jensfelt. Active visual object search in unknown environments using uncertain semantics. *IEEE Transactions on Robotics*, 29(4):986–1002, 2013.

[4] A. Aydemir, K. Sjöö, J. Folkesson, A. Pronobis, and P. Jensfelt. Search in the real world: Active visual object search based on spatial relations. *2011 IEEE International Conference on Robotics and Automation*, pages 2818–2824, 2011.

[5] D. H. Ballard. Active Perception. *Encyclopedia of Neuroscience*, (March):31–37, 2009.

[6] S. Clippingdale and R. Wilson. Self-similar Neural Networks Based on a Kohonen Learning Rule. *Neural Networks*, 9(5):747–763, jul 1996.

[7] R. P. de Figueiredo, A. Bernardino, J. Santos-Victor, and H. Araújo. On the advantages of foveal mechanisms for active stereo systems in visual search tasks. *Autonomous Robots*, 42(2):459–476, 2018.

[8] M. Grotz, T. Habra, R. Ronsse, and T. Asfour. Autonomous view selection and gaze stabilization for humanoid robots. In *2017 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 1427–1434, 2017.

[9] L. M. Kaplan, S. Chakraborty, and C. Bisdikian. Fusion of classifiers: A subjective logic perspective. In *2012 IEEE Aerospace Conference*, pages 1–13, 2012.

[10] J. H. Krantz. The Stimulus and Anatomy of the Visual System. In *Experiencing Sensation and Perception*, chapter 3, pages 3.1 – 3.36. 2012.

[11] C. Melício, R. Figueiredo, A. F. Almeida, A. Bernardino, and J. Santos-Victor. Object detection and localization with Artificial Foveal Visual Attention. In *2018 Joint IEEE 8th International Conference on Development and Learning and Epigenetic Robotics (ICDL-EpiRob)*, pages 101–106, 2018.

[12] T. Minka. Estimating a Dirichlet distribution. *Technical report, M.I.T.*, 2000.

[13] L. Montesano and M. Lopes. Learning grasping affordances from local visual descriptors. In *2009 IEEE 8th International Conference on Development and Learning*, pages 1–6, 2009.

[14] J. Redmon and A. Farhadi. YOLOv3: An Incremental Improvement, 2018.

[15] B. Settles. Active Learning Literature Survey. Computer Sciences Technical Report 1648, University of Wisconsin–Madison, 2009.

[16] J. P. Siebert, P. Ozimek, L. Balog, N. Hristozova, and G. Aragon-Camarasa. Smart Visual Sensing Using a Software Retina Model. In *IROS2018 Workshop: Unconventional Sensing and Processing for Robotic Visual Perception, at 2018 IEEE/RSJ International Conference on Intelligent Robots and Systems*, 2018.