TÉCNICO LISBOA

Semantic-Based Active Perception for Humanoid Visual Tasks

Scene Exploration and Visual Search in Foveal Scenes using Deep Object Detection Models

João Miguel Barradas Luzio

Thesis to obtain the Master of Science Degree in

Electrical and Computer Engineering

Supervisors: Prof. Alexandre José Malheiro Bernardino
Prof. Plinio Moreno López

Examination Committee

Chairperson: Prof. João Manuel de Freitas Xavier
Supervisor: Prof. Alexandre José Malheiro Bernardino
Member of the Committee: Prof. Rui Pimentel de Figueiredo

November 2023

Declaration

I declare that this document is an original work of my own authorship and that it fulfills all the requirements of the Code of Conduct and Good Practices of the Universidade de Lisboa.

Acknowledgments

Believing that some of the most precious treasures that lie at the end of my academic journey are the memories that accompany the path I have traveled together with those with whom I have forged a great bond of friendship along the way, I would like to thank my colleagues Gonçalo Gomes, Hugo Matias, and Pedro Taborda for their companionship throughout the years. Furthermore, I would especially like to thank my working partner José Reis, for you have been by my side since day one, collaborating and supporting me. I shall never forget that.

Secondly, I would like to thank my supervisors, Prof. Alexandre Bernardino and Prof. Plinio Moreno, for all the great advice, the knowledge shared, and the many interesting conversations during the development of this Thesis. I am extremely grateful for your effort and guidance during this challenge. I also would like to extend my gratitude to Pedro Vicente, Ph.D., for helping me figure out my path.

Thirdly, I want to express my deepest gratitude to my beloved parents, my lovely sisters, my brothers-in-law, and my nephews, for their love, support, and heartwarming presence in my life. Thank you my dear friends Ana Isabel, Denise, and Jonas for always cheering me up and to everyone else who has somehow contributed to my success. You are all very special indeed. I would also like to express my deepest appreciation to my lovely fiancee. Your overwhelming love has been truly manifested in my life. Since you have been there from the beginning, nothing feels so special as sharing this personal accomplishment with you. Moreover, I extend my sincere gratitude to my fiancee's family, for all the motivation and incentive, for listening to my concerns, and for always giving me your best advice.

Finally, I praise and give glory to the One through whom all things are made new, the Alpha and the Omega, whose unrevealed mysteries are the motive behind my interest and dedication to science itself.

"Men became scientific because they expected Law in Nature, and they expected Law in Nature because they believed in a Legislator."

– C. S. Lewis, *Miracles*

Abstract

In this work, it is studied how well a semantic-based foveal active perception model is able to complete visual tasks that are regularly performed by humans, namely, scene exploration and visual search. To this end, the accuracy of the semantic-based model is compared with with the accuracy of a traditional saliency-based model, derived from past developments in the fields of neurology and psychology, that attempt to reproduce aspects of humanoid visual cognition. Regarding the scene exploration task, the semantic-based approach convincingly outperforms the traditional saliency-based model, when it comes to accurately mapping the semantic content contained by the visual field. In light of the visual search experiments, it was concluded that the semantic-based predictive approach significantly outperforms the saliency-based model, as well as a random gaze selection algorithm, both in accuracy and relative computational cost. The latter results were obtained while searching for instances of a target class in a visual field containing multiple distractors.

Keywords

Active Perception; Foveal Vision; Object Detection; Scene Exploration; Visual Search.

Resumo

Neste trabalho, estuda-se até que ponto um modelo de perceção ativa foveal baseado, baseado em informação semântica, é capaz de completar tarefas visuais que são regularmente praticadas por humanos, nomeadamente, exploração do campo visual e a procura visual de objetos. Para o efeito, a precisão do modelo baseado em informação semântica é comparada com a de um modelo tradicional baseado em análise de saliência, derivado de desenvolvimentos nos domínios da neurologia e da psicologia, que tenta reproduzir aspetos da cognição visual humana. Relativamente à tarefa de exploração de cenários visuais, a abordagem baseado em informação semântica supera de forma convincente o modelo tradicional, baseado em saliência, quando se trata de mapear com precisão o conteúdo semântico contido no campo visual. À luz das experiências de procura visual, concluiu-se que uma abordagem preditiva em informação semântica supera significativamente o modelo de saliência, bem como uma seleção aleatória do próximo ponto focal, tanto em termos de precisão como de custo computacional relativo. Estes últimos resultados foram obtidos durante a procura de instâncias de uma classe-alvo num campo visual, na presença de múltiplos objetos de distração.

Palavras Chave

Perceção Ativa; Visão Foveal; Deteção de Objectos; Exploração do Campo Visual; Procura Visual.

Contents

List of Figures

List of Tables

Acronyms

AET	Attentional Engagement Theory
AP	Average Precision
AUC	Area Under Curve
CNN	Convolutional Neural Network
COCO	Common Objects in Context
FIT	Feature Integration Theory
GRF	Gaussian Receptive Field
iNVT	iLab Neuromorphic Vision Toolkit
IoU	Intersection over Union
IOR	Inhibition of Return
IVSN	Invariant Visual Search Network
KL	Kullback-Leibler
mAP	mean Average Precision
MLE	Maximum Likelihood Estimation
NMS	Non-Maximum Suppression
PDF	Probability Density Function
RoI	Region of Interest
R-CNN	Regional Convolutional Neural Network
SEM	Standard Error of the Mean
SSD	Single-Shot Detector
WTA	Winner-Take-All
YOLO	You Only Look Once

1

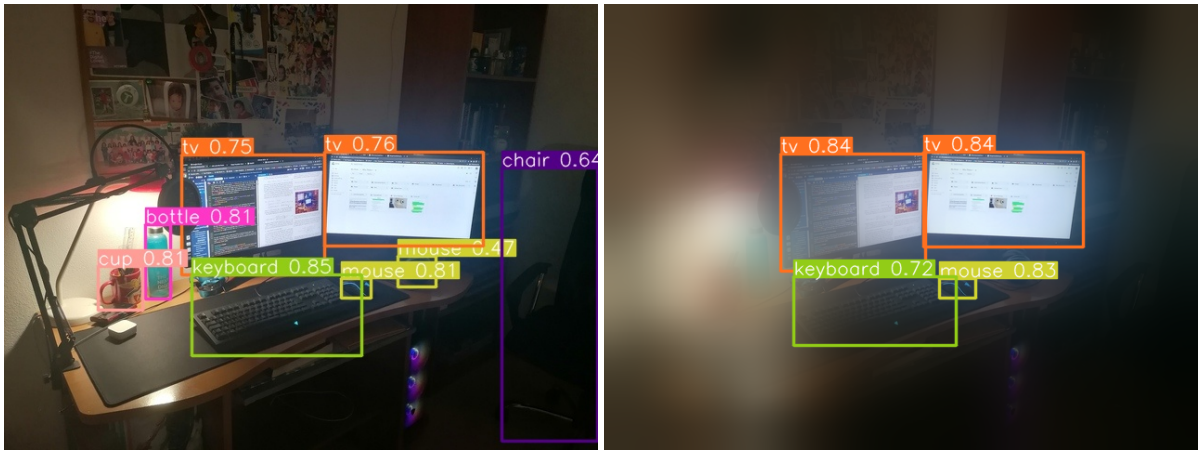
Introduction

Contents

As pointed out by Bajcsy et al. [4], the topic of perception, and more specifically visual perception, has been a great topic of discussion through the years within the scientific community. This work revolves around a combination of active perception and semantic information, provided by modern object detection models [18], to complete visual tasks that are frequently performed by humans. Furthermore, to properly emulate the usage of sensors [19], whose structure is inspired by the human eye, it is considered the integration of an adequate visual system, that permits the extraction of intricate information from the environment. Chapter 1 introduces the work developed in the context of this dissertation.

1.1 Motivation

Active perception is a cognitive process that involves actively gathering information from the environment, through various sensors, and then using this information to make decisions or take actions [4]. More specifically, active visual perception [4,19] especially focuses on the way both biological organisms and artificial systems actively interact with the environment to extract relevant information. This can involve the performance of actions such as actively moving sensors (e.g. eyes, in a biological context [5], and cameras, on artificial systems [20]) to change the viewpoint or modifying visual parameters to gather specific information. This concept is particularly important in robotics, computer vision, and fields where machines or systems need to interact with their surroundings in an intelligent and adaptive manner [4].



(a) Regular Cartesian Image

(b) Artificially Foveated Image

Figure 1.1: A regular image (a) that has been artificially foveated (b) to emulate a visual field, together with the respective You Only Look Once (YOLO)v3 [1] object detections, represented by their bounding-boxes.

The human visual system is a complex sensory apparatus that allows humans to perceive and interpret the surrounding environment. It comprises several key components [21]: the eyes, which capture light and convert it into neural signals, the optic nerves, which transmit these signals to the brain, and

the brain itself, which processes and interprets the visual information. The eyes have specialized structures [6] like the cornea and lens that focus light onto the retina, which is a light-sensitive layer at the back of the eye containing photoreceptor cells called rods and cones. Rods are responsible for low-light vision, while cones enable color discrimination and fine detail perception. The central region of the retina, known as the fovea [5], is densely populated with cones and provides the sharpest visual acuity, as corroborated by the artificial visual field presented in Figure 1.1(b). From the retina, signals are sent through the optic nerves to the brain's visual processing centers, particularly the primary visual cortex. Here, the brain assembles the signals into a coherent visual perception, allowing humans to recognize shapes, colors, motion, and depth [6]. This intricate system enables human beings to navigate and interact with the environment, making vision one of our most vital senses. A foveal computational system [8, 19] is able to reduce the amount of information to be cognitively processed in each gaze fixation [5]. Therefore, integrating all the chunks of information collected along the visual exploratory path into a short-term memory structure appears to be of great advantage when it comes to minimizing computational costs, as the agent does not need to store and process all the information at once [22].

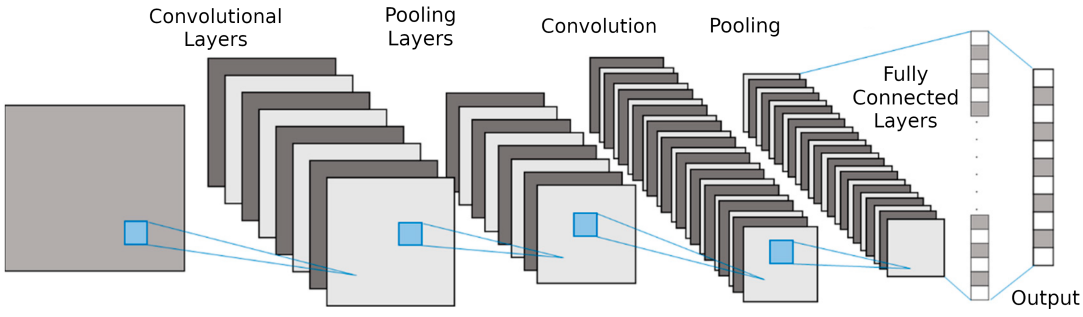


Figure 1.2: Illustration of the general architecture of a Convolutional Neural Network (CNN) [2].

As illustrated in Figure 1.2, CNNs are a class of deep learning models [2] specifically designed for processing data with a grid format. CNNs excel in feature extraction through the application of convolutional operations, where filters slide across the input data, performing element-wise multiplication and summation operations [10, 23, 24]. This allows the respective neural network to learn hierarchical representations, starting from basic patterns and gradually building up to more complex features. Pooling layers further enhance efficiency by reducing spatial dimensions while retaining significant information. Fully connected layers integrate these features for final predictions or classifications. CNNs have demonstrated remarkable efficacy in computer vision tasks [19, 22], leveraging their capacity to extract relevant features. This hierarchical approach is particularly effective in tasks such as object detection [18], where objects can be characterized by a combination of low-level and high-level features. In recent years, modern deep object detection models [18] have been taking advantage of the overwhelming ability of CNNs to extract relevant hierarchical features from visual representations to localize and classify objects.

In the context of humanoid robotics [19], recently developed visual-cognitive models (e.g. [3, 25]) consider the incorporation of semantic content, extracted by modern deep object detection models [18], to perform visual tasks, such as scene exploration [26], visual target search [17], and environment recognition. This work dives deep into this matter, assessing whether an already proposed semantic-based methodology [3] is able to accurately complete a couple of well-known humanoid visual tasks.

1.2 Objectives

Traditional cognitive models from psychology [19] state that the guiding mechanism for human visual attention depends on a range of distinct types of conspicuous features [27], which are combined in a master map that highlights potential regions of interest. For this reason, modern cognitive approaches consider visual processing models that are able to extract either bottom-up [9, 15], top-down [22, 28], or both types of features [29] to actively perform common human visual tasks. A recent scientific study developed by Dias et al. [3], regarding humanoid visual perception [4], explores the possibility of combining the rich semantic information provided by modern object detectors [18] with active perception, to perform scene exploration. When exploring a scene, the goal is set on accurately mapping the semantic content of a visual field through successive premeditated gaze shifts, governed by a proper mechanism.

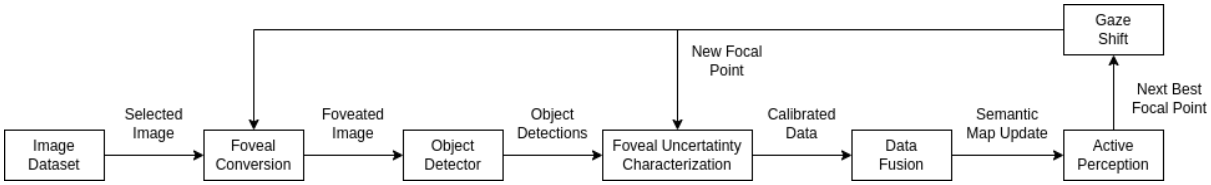


Figure 1.3: Full methodological pipeline, as proposed by Dias et al. [3], for active perception with foveal vision [4,5].

Despite providing, implementing, and testing the complete methodological pipeline used to perform scene exploration, displayed in Figure 1.3, Dias et al. [3] does not fully establish how successful the proposed model is in achieving the main goal of the task. This central goal consists of adequately mapping the maximum amount of semantic content available in the visual field. Furthermore, it is also considered the adaptation of the same methodology [3] to another visual task that is commonly performed by humans, namely visual search [17], where the objective is to set the gaze directly upon a region that contains at least one instance of a pre-defined target class. For this reason, the relevance of the gaze selection mechanism is amplified, given the greedy target-oriented nature of this cognition-related task.

In summary, the objective of this work is to establish how well semantic-based active visual perception is able to perform the described visual tasks, namely scene exploration [26] and visual search [17]. For this purpose, the accuracy of a semantic model [3] is compared with the accuracy of a reformed biologically inspired model [15], which serves as a benchmark for human visual cognition. This baseline

model [9] attempts to mimic both the behavior and neuronal architecture of the primate visual system [9]. Furthermore, the aim is to confirm whether the adopted semantic-based model [3] is applicable in a visual search context [17], by taking advantage of the images contained in the renowned Common Objects in Context (COCO) dataset [16], which can be foveated with the aid of an artificial visual system [8] to simulate the intricacies of a humanoid visual field. In the context of visual search, it is assumed that a simulated visual field must contain at least one object of interest and multiple distractors.

1.3 Organization of the Document

The outline of this document is the following: In Chapter 2 is presented a compilation of key concepts, theories, and topics that are foundational to this work [19], such as human visual cognition, object detection, and active perception. Then, Chapter 3 describes the full methodological apparatus that involves semantic-based foveal active perception [3], which will be applied during the experimental phase. Continuing the sequence, in Chapter 4, the full experimental apparatus is meticulously described, together with the respective test conditions and constraints. Chapter 5 presents all the relevant results obtained during the experimental phase, upon completing both scene exploration [26] and visual search [17] tasks, using different active perception methods [3, 15]. Finally, in Chapter 6, important contributions are highlighted, and experimental remarks and conclusions are presented, together with thoughtful considerations about future scientific endeavors that may proceed from the work presented in this document.

2

Background & Related Work

Contents

In this chapter, there are presented key concepts and theories that are essential when aiming to comprehend the methodology that will be applied in this work to its fullness. Moreover, there is an in-depth compilation of topics that are foundational to the fields of visual cognition and active perception.

2.1 Humanoid Visual Cognition

Foveal vision [5] and active perception [4] are the bedrock of both human visual and cognitive systems [19]. Foveal vision reduces the amount of information to process during each gaze fixation, while active perception changes the gaze direction to the most promising regions of the visual field. In a similar fashion, humanoid robots can also be able to explore a scene [25], identifying objects displaced in their surrounding environment, if properly equipped with adequate sensors and cognitive emulation software.

2.1.1 Foveal Vision

Despite being often compared to a photographic camera, the human eye's processing capacity across the visual field is not homogeneous. Several anatomical properties of the eye are correlated with the presence of gaps [5] in sensory information. Due to these anatomical properties, the processing of visual signals varies quite dramatically across the visual field. Therefore, it is important to distinguish between the center of the visual field, known as the fovea, and an outer region, known as the periphery.

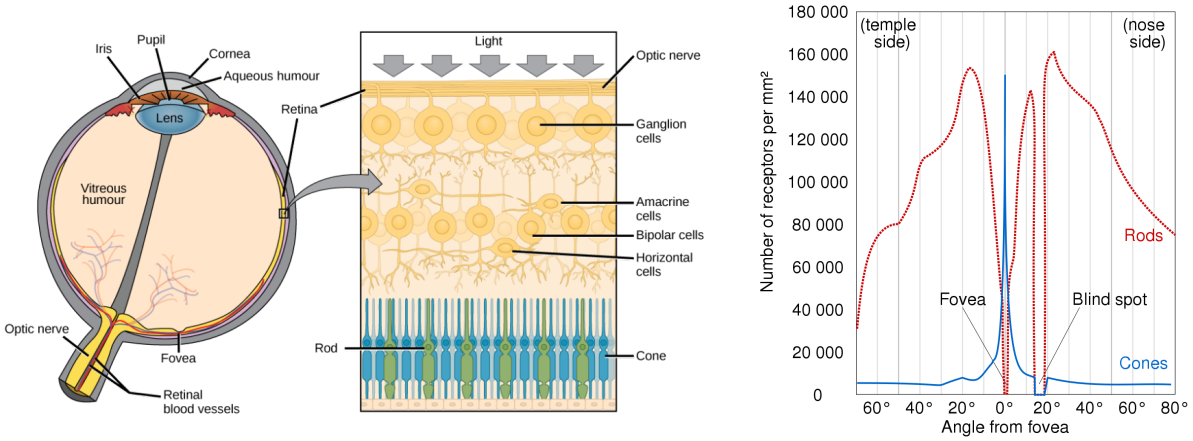


Figure 2.1: Illustration of the anatomy of the human eye together with a representation of the retinal layers, as well as a graphical depiction of the density distribution of cone and rode cells [6] across the full foveal range.

Several anatomical properties of the eye are correlated with the presence of gaps in sensory information [5]. In the human eye's retina reside two fundamental types of photoreceptors [6], namely the cone and rod cells. Essentially, there are three types of cone photoreceptors, each being able to distinguish one color out of the red, green, and blue colors. The rod photoreceptors sense brightness contrasts, being prone to detect motion in the surrounding environment. Furthermore, as illustrated in Figure 2.1

there are no photoreceptors at the optic disk [6], where the ganglion cell axons exit the retina. Hence, a so-called blind spot is found in that region of the retina. Due to these anatomical properties, the processing of visual signals varies quite dramatically across the visual field. Foveal vision allows for maximum acuity and contrast sensitivity in a small area around the gaze position, while peripheral vision allows for a large field of view [5], although it presents lower resolution and contrast sensitivity, as well as higher positional uncertainty and crowding. Visual crowding is a perceptual phenomenon [30] where an object that is recognized and identified in a location far away from the fovea, i.e. in the peripheries, is impaired by the presence of other neighboring objects, due to a texture-processing neurological mechanism.

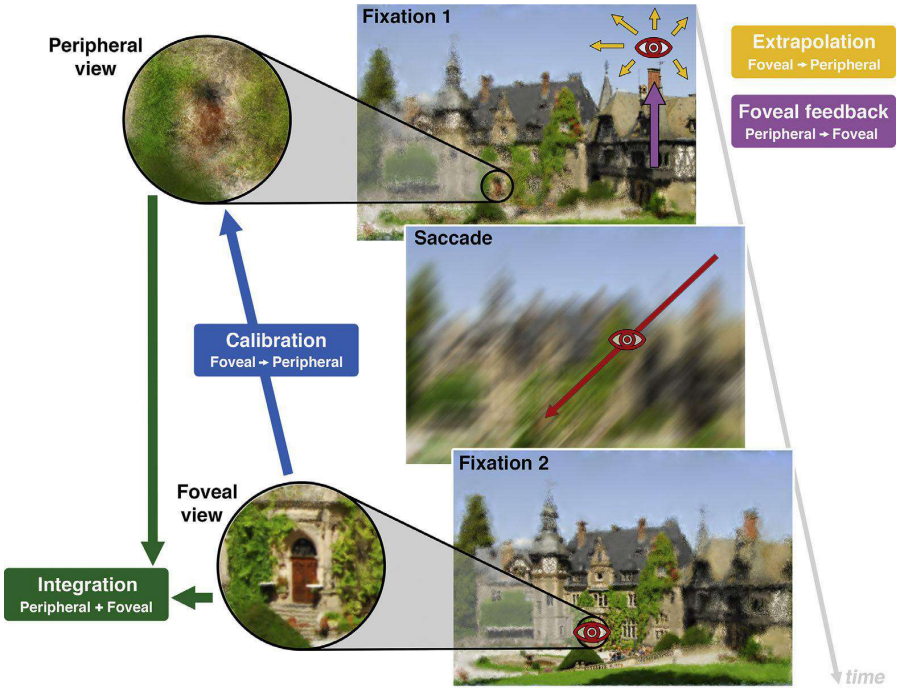
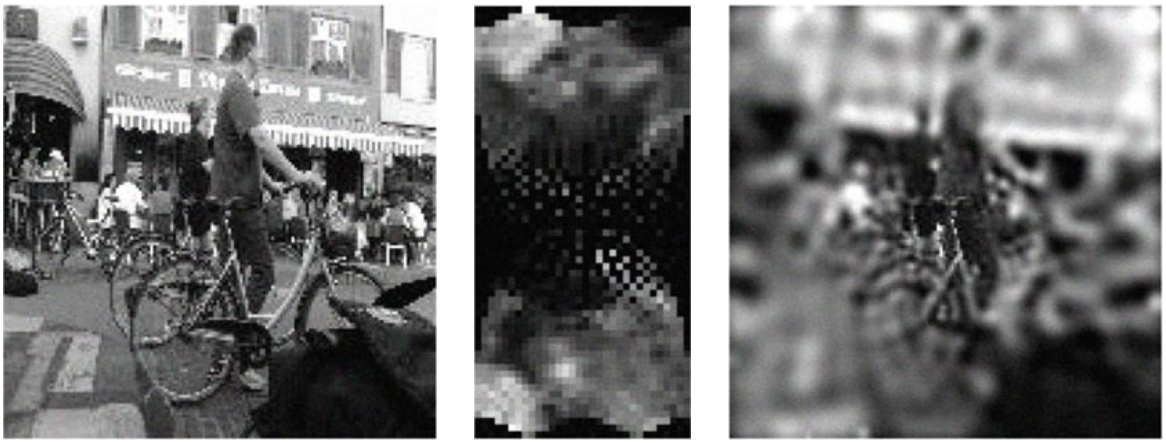


Figure 2.2: Schematic illustration [5] of interactions between peripheral and foveal vision. The images illustrate differences between peripheral and foveal vision and the typical sequence of transsaccadic vision [6]. Yellow, purple, green, and blue arrows indicate the direction of information flow during the saccade.

A saccade [5] is a conjugate rapid eye movement that shifts the center of gaze from one part of the visual field to another. These fast and instinctive eye movements, which are schematically illustrated in Figure 2.2, are mainly used to guide the gaze toward objects of interest. Saccades allow the eyes to move rapidly toward visual, auditory, or tactile stimuli and contribute to the identification of points of interest in the surrounding environment [6] when performing any common human cognitive task. the performance of sequential premeditated saccades is part of the backbone of the active visual perception [4], especially when considering a move to a new fixation point. Humanoid agents tend to plan their saccades with a specific objective in mind, whether it involves reaching a portion of the visual field otherwise hidden due to occlusion, observing a larger portion of the surrounding visual world, or compensating for spatial non-uniformity of the visual system [5], such as in the context of foveal vision.

2.1.2 Image Foveation Techniques

The representation within the human brain diverges from the visual input received by the retina. Remarkably, the brain of a human possesses the extraordinary ability to seamlessly reconstruct and autonomously compensate for any disparities between these two perceptual streams [20,27]. This intricate process ensures that the human's perception of the world remains coherent and consistent, despite the underlying variances in the raw visual data [21]. A classical approach to generate foveated images from regular Cartesian images is to use methods based on log-polar transformations to obtain cortical maps [31]. A cortical map is a topographic representation of sensory or motor information in the brain, specifically found in the cerebral cortex. A log-polar space is the most appropriate approach to generate cortical images since it has been shown to model with reasonable fidelity the mapping observed in the primate visual cortex [32]. Mathematically, this mapping provides a plausible model that provides the key geometric features of the fovea [5,7] and the compression of the peripheral visual field, accordingly.



(a) Original Image **(b)** Cortical Map **(c)** Foveal Reconstruction

Figure 2.3: Example of a regular Cartesian image (a) and its respective cortical map (b), from where the original image can be reconstructed into a foveal image (c), using adequate foveal reconstruction methods [7].

Log-polar cortical mapping approaches [31,32] use log-polar coordinates, more specifically θ and ρ, corresponding to the angle and the logarithmic distance from the central point of the fovea, respectively

$$\rho = \log\left(\sqrt{x^2 + y^2}\right) \quad \text{and} \quad \theta = \operatorname{atan2}(y, x)\ , \tag{2.1}$$

where variables x and y represent the Cartesian coordinates relative to the origin of any given image. The log-polar space suffers from severe sparsity in the foveal region and excessive density at the peripheries. This has been mitigated by Ozimek et al. [32] by removing the log operator from (2.1) and switching to the quasi-linear polar space. As pointed out by Traver and Bernardino [31], another weakness of this

model is the existence of a singularity in the center of the image. The most common approach to solve this problem consists of adding an α parameter to the model, where the new coordinates are defined as

$$\mathrm{X}_{\text{cort}} = \sqrt{(x+\alpha)^2 + y^2} \quad \text{and} \quad \mathrm{Y}_{\text{cort}} = \operatorname{atan2}(y, x+\alpha) \ , \tag{2.2}$$

setting the field of view of the quasi-linear region of the retino-cortical polar mapping [21]. There are several classic methods [31] to reconstruct a foveated image (also known as a back-projection) from the generated cortical map, e.g. Super-Pixels and Gaussian Receptive Fields (GRFs), as in Figure 2.3. To perform retinal sampling accurately and efficiently, Ozimek et al. [32], inspired by Pamplona and Bernardino [7], applies GRFs, which follow the stochastic nature and the biological retinal architecture.

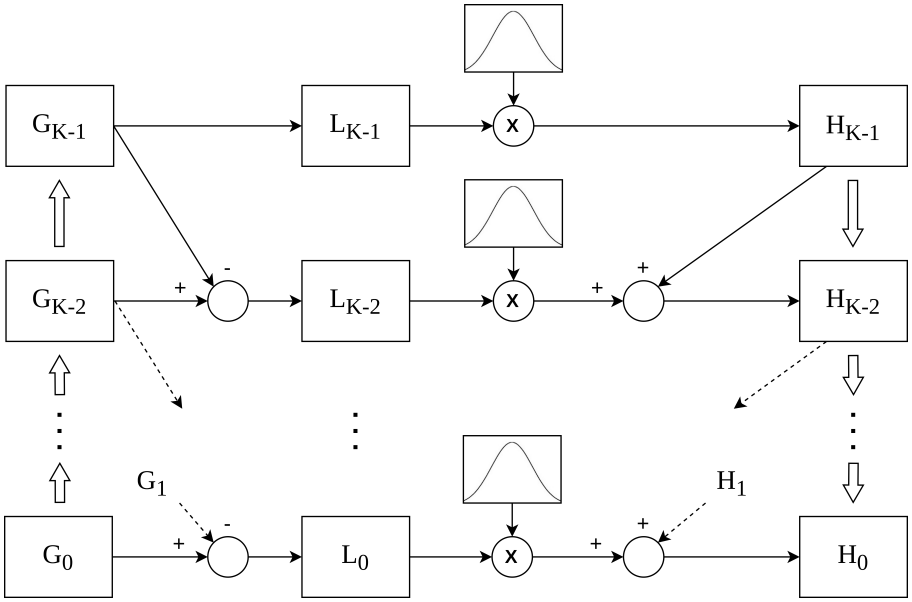


Figure 2.4: Schema of the artificial foveal system [8] that summarizes the steps in a foveation system comprised by K levels. The Gaussian pyramid level G_0 corresponds to the original image and H_0 to the foveated image. The thick up arrows represent sub-sampling and the thick down arrows represent up-sampling.

After generating the back-projected images, Siebert et al. [32] consider the usage of a CNN, known as the 4196 Node Retina, to classify any object that may be present in a specific fixation. The tests were carried out using the ImageNet dataset [33], a popular image dataset commonly used in the context of object detection. Furthermore, other cortical mapping approaches, such as Lukanov et al. [34], even consider applying Cartesian foveal geometry to avoid the nonlinearities [31] associated with log-polar spaces. Unlike methods based on log-polar transformations, which are complex to implement, the method proposed by Almeida et al. [8,22] is applicable in real-time image processing, which is extremely important when considering human-like visual systems for visual tasks. The schema presented in Figure 2.4 corresponds to the artificial foveal system [8], which is inspired by a Laplacian pyramid technique [35] for image compression, replicating the receptive fields that are present in the eyes of a human being [5].

2.1.3 Artificial Foveal System

The artificial foveal system [8], schematized in Figure 2.4, consists of four steps. The first step involves creating a Gaussian pyramid composed of K levels. This is essentially a Gaussian scale space, where each level is a low-pass-filtered version of the preceding level. Each level increases the amount of blur and is generated based on the preceding level. To obtain G_k, that is, the image at level k, the image G_{k-1} is convoluted with a two-dimensional isotropic and separable Gaussian filter kernel, defined as

$$g_k\left(u, v, \sigma_k\right) = \frac{1}{2\pi\sigma_k}\, e^{-\frac{u^2+v^2}{2\sigma_k^2}}, \quad 0 < k < K\,, \tag{2.3}$$

and scaled down, where u and v represent the coordinates of the pixels in the image and $\sigma_k = 2^{k-1}\sigma_1$ represents the standard deviation at level k. The second step consists of up-sampling (interpolating) all the G_k images to impose the same resolution at all levels. For the next step, a Laplacian pyramid [35] is computed from the difference between adjacent Gaussian levels. The Laplacian pyramid consists of a series of error images, where each level represents the difference between two levels of the previous output. Finally, the last step consists of multiplying exponential weighting kernels, which are given by

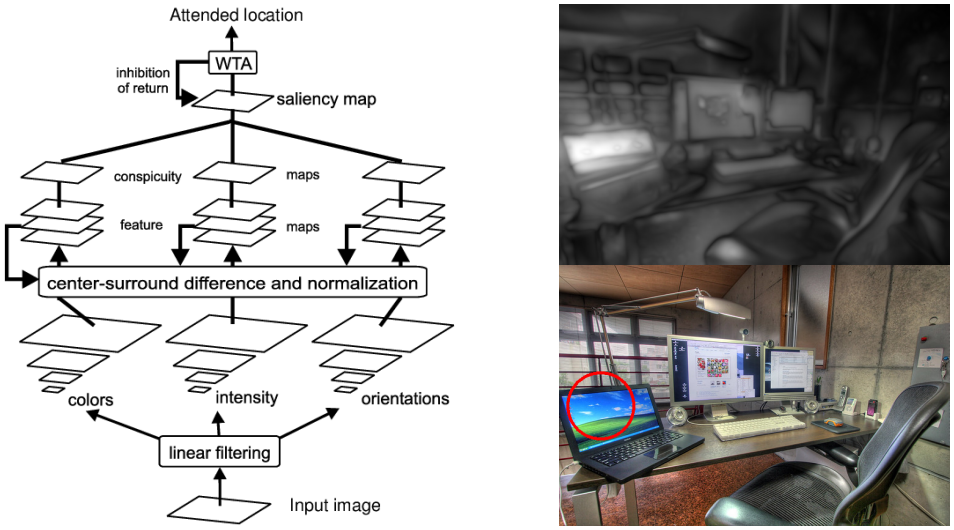
$$h_k\left(u, v, f_k\right) = e^{-\frac{(u-u_0)^2+(v-v_0)^2}{2f_k^2}}, \quad 0 \leq k < K\,, \tag{2.4}$$

with each level of the Laplacian pyramid to emulate a smooth fovea. The size of the region with the highest acuity, known as the foveal size and defined by f_0, so that $f_k = 2^k f_0$ expresses the exponential kernel standard deviation at the level k. The focal point of attention, called the foveation point, has coordinates (u_0, v_0). The choice of the focal point is of the utmost importance when considering the context of active perception [4, 19], since, to extract relevant information, critical portions of the image should be contained in the nitid foveal region of the visual field [3], corresponding to the central point.

2.1.4 Cognitive Attention Mechanisms

Biological organisms use selective visual attention mechanisms [9] to process single portions of the available visual information while disregarding the remainder. This enables these organisms to efficiently perceive and handle the limited neural resources, in the brain, which are allocated to vision. One of the most important and challenging tasks in computer vision is visual saliency detection, which aims to highlight the most dominant object regions within an image, also known as Region of Interests (RoIs). Numerous applications incorporate visual saliency to improve their performance [17, 36], including object detection [18]. Moreover, visual attention covers all factors that influence the information selection mechanisms [19], whether guided by visual stimuli, which are referred to as bottom-up features, or by task-related expectations, denominated as top-down features. Bottom-up approaches rely on local fea-

ture contrasts [9, 37], extracted from the visual field. However, this low-level feature extraction method is limited in its ability to consider multi-scale and high-level semantic information, leading to the generation of low-contrast salient maps rather than actual salient objects [38]. On the other hand, top-down approaches use object categories to generate salient maps, often through semantic segmentation [28].



(a) Attention Mechanism's Schema **(b)** Saliency Map

Figure 2.5: Itti-Koch [9] visual attention system (a), inspired by the neural architecture of the primate visual system, together with an example of a saliency map (b), with the respective most salient region (red circle).

The Feature Integration Theory (FIT) [17] states that in the human brain, several features are processed in parallel in different areas of the brain, and the locations of these features are collected in a single master map of locations. From this map, an attention-based mechanism selects the most promising RoI. This master map is similar to what is nowadays called a saliency map, and there is strong evidence [39] that this map exists in the brain. The Attentional Engagement Theory (AET), also known as the similarity theory, asserts that stimulus objects are represented at a level of perceptual description, where top-down object representations [17, 36] compete with each other to enter the visual short-term memory. A classical neural architecture proposed by Itti et al. [9], schematized in Figure 2.5, considers a pure bottom-up approach to generate a saliency master map that is constructed on the posterior parietal cortex of primates. Following the principles established in the FIT [17], this architecture takes advantage of extracted positional features, considering intensity and color contrasts to generate multiple layered feature maps, which represent center-surround contrasts that are typically present on RoIs. Finally, following the diagram presented in Figure 2.5(a), a Winner-Take-All (WTA) selection mechanism identifies the most promising RoI out of a master map that combines conspicuity maps, in which different feature layers are fused. Also, an Inhibition of Return (IOR) mechanism [9, 29] ensures that, when transversing the field of view to complete visual tasks, previously visited RoIs are ignored by the WTA mechanism.

2.2 Semantic Object Detection

Object detection models aim at classifying existing objects in any provided image, localized inside the boundaries of a rectangular box (bounding-box) with an associated confidence score [19]. There are three main stages in traditional object detection models [18]: selection of informative regions, extraction of features, and classification of objects. Several approaches follow this pipeline when solving object detection, each of them unique in some particular aspect. Object detection frameworks can mainly be divided into two types: The first approach (two-stage frameworks) follows the traditional pipeline, consisting of generating region proposals first and then classifying each proposal into the relevant object categories (e.g. Regional Convolutional Neural Network (R-CNN) [24]). In the second approach (one-stage frameworks), object detection is viewed as a regression or classification problem, where both the categories and locations are determined directly through a unified framework (e.g. Single-Shot Detector (SSD) [23], YOLO [10]). These frameworks have been propelled by the impressive power of Convolutional Neural Networks (CNNs) [2], given their ability to process hierarchically extracted features.

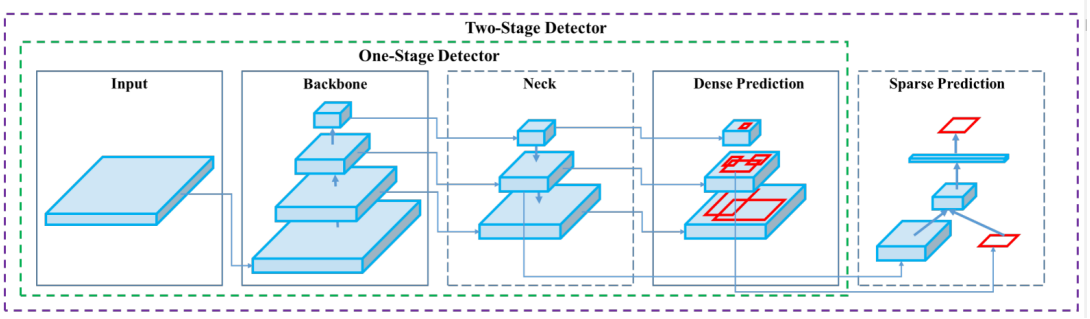


Figure 2.6: Schematic presentation of the YOLO [10, 11] framework, representing the pipeline for the RoI identification mechanism, starting from the input image. For instance, YOLOv3 uses the Darknet-53 [12] as its backbone CNN architecture [2]. This schema is originally presented in the YOLOv4 documentation [13].

For instance, the You Only Look Once (YOLO) framework, schematized in Figure 2.6, was proposed by Redmon et al. [10] as a design that makes use of the entire top-most feature map to predict both confidences for multiple categories and bounding-boxes [11]. The basic idea behind the YOLO [10] model is to divide the input image into a grid, and each grid cell is responsible for predicting the object that is centered on it, establishing its respective bounding-box and corresponding confidence scores. Finally, the resulting detections are filtered with a Non-Maximum Suppression (NMS) algorithm [40], a technique used to suppress overlapping bounding-boxes, eliminating those with low confidence scores, based on a pre-defined overlap [14] threshold. Moreover, recent versions of YOLO, such as YOLOv3 [1], adopt several improvement strategies, such as batch normalization, which consists of normalizing the output of a layer by subtracting the mean and dividing by the standard deviation of the output of the layer over a mini-batch. It also adopts pre-defined anchor boxes [41], which are essentially default bounding-box guesses that are applied to speed up the RoI prediction process [1], and multi-scale image training.

2.3 Evaluation Metrics

Regarding object detection, evaluation metrics permit assessing both the RoI localization precision and classification accuracy. Furthermore, when considering any objective-driven iterative task, it is also crucial to measure how close is an agent to achieving the goal, in order to properly evaluate its performance.

2.3.1 Object Detection Metrics

The mean Average Precision (mAP) is the most common metric used in Deep Learning [42] to evaluate the robustness of object detection models. To better understand this performance measurement it is vital to first understand the concept of Intersection over Union (IoU) [14], also known as *Jaccard* index.

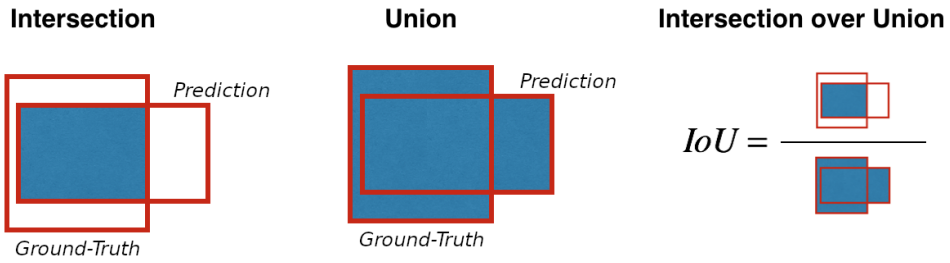


Figure 2.7: Graphical representation of the *Jaccard* index [14] between ground-truth and predicted bounding-boxes.

The purpose of IoU is to indicate the overlap of the predicted bounding-box coordinates with the ground-truth box of an identified object, as visually explained in Figure 2.7. The higher the IoU score, the closer the predicted bounding-box coordinates are to the ground-truth box coordinates. By defining an IoU threshold, it is then possible to classify the predicted bounding-boxes as true positives (TP), i.e. the model's prediction correctly matches the ground-truth, or false positives (FP), i.e. the model's prediction does not coincide with the ground-truth. From these concepts emerge two distinct metrics:

$$\text{Precision} = \frac{\mathrm{TP}}{\mathrm{TP}+\mathrm{FP}} \quad \text{and} \quad \text{Recall} = \frac{\mathrm{TP}}{\mathrm{TP}+\mathrm{FN}}\,. \tag{2.5}$$

Precision measures how well one can find true positives (TP) from all positive predictions ($\mathrm{TP}+\mathrm{FP}$). A false negative (FN) occurs when the model does not predict a label, but it is part of the ground truth. Recall measures how well one can find true positives (TP) from all predictions ($\mathrm{TP}+\mathrm{FN}$). The precision of a model indicates how reliable its positive predictions are [14], whereas the recall of a model indicates whether it missed any predictions that it should have made. Over the years, researchers have tried to combine precision and recall into a single metric to compare object detection models [43]. There are a couple of metrics, that combine precision and recall, that are widely used, such as the F1-score [44] and the Average Precision (AP) [42]. In order to fully comprehend the mAP, one must understand the AP.

The AP is a metric that allows for comparison between the performance of object detection models. Given a specific IoU threshold, the predictions can be ordered according to some prediction confidence score. Then, precision and recall (2.5) can be calculated for different confidence scores, in an interval of N values that systematically cover the entire range, allowing for the plot of a discrete function $p(r)$, which represents precision as a function of recall [43]. The AP [42] is computed through the expression

$$\text{AP} = \frac{1}{N} \sum_{r} p(r)\,. \tag{2.6}$$

which is the average value of $p(r)$ over the interval of N values. Precision-recall curves, which typically consist of interpolating the function $p(r)$, not only encapsulate the trade-offs of both metrics [43] but also maximize their effects, as it better accounts for the model's overall accuracy. Due to the overall coverage of the area, the precision-recall Area Under Curve (AUC), which is an alternative metric computed as

$$\text{PR-AUC} = \int_{0}^{1} p(r)\, dr \tag{2.7}$$

where $p(r)$ is an interpolation of the N points, is considered a superior metric when compared to the AP [42]. Therefore, the PR-AUC consists in finding the area underneath the precision-recall curve. The mAP metric incorporates the trade-off between precision and recall and takes into account both false positives (FP) and false negatives (FN). This property makes the mAP [42], computed for K classes as

$$\text{mAP} = \frac{1}{K} \sum_{n=1}^{K} \text{AP}_n\,, \tag{2.8}$$

a suitable metric for most detection applications. The calculation of mAP varies depending on certain specifications. For instance, mAP can also be calculated by averaging the AP not only over K object classes but also over several IoU thresholds. For instance, the IoU can vary from 0.5 to 0.95 with a step size of 0.05, as in the famous COCO mAP metric [16]. Ultimately, calculating mAP over an IoU threshold range avoids the ambiguity of choosing the optimal IoU threshold for the evaluation of a detection model.

2.3.2 Task-Specific Metrics

In this work, the performances of different active visual perception [4] methods are assessed, with respect to the completion of humanoid visual tasks. For this reason, it is also important to discuss what are the benchmark evaluation metrics that are suitable to the tasks at hand. Regarding the scene exploration task [3], since the main objective is to accurately map the semantic content of the scene, which involves finding and correctly identifying the maximum number of objects displayed in the visual field, it would be interesting to consider precision-recall inspired metrics (2.5). For instance, a recall-like met-

ric promotes a global perspective on how much correct information has been gathered (TP) in view of the full extent of ground-truth ($\mathrm{TP} + \mathrm{FN}$) information. Regarding visual search [17], the proponents of Invariant Visual Search Network (IVSN) [29] compare the performances of different approaches through the cumulative performance. Essentially, the cumulative performance assesses how many instances of the test set have been completed (i.e. the target has been effectively found) at a given iteration. Other metrics measure the distance either between saccades or the focal point and a target object [29], in an evolutive perspective, assessing the variation across multiple saccades, upon the task's completion.

2.4 Principles from Probability Theory

This section consists of an exposition of fundamental probabilistic concepts [45] that serve as the baseline for the methodology. When manipulating semantic content, one must understand concepts such as categorical and Dirichlet distributions [46], used in Bayesian inference for calibration processes [47, 48].

2.4.1 Categorical and Dirichlet distributions

A categorical probability distribution characterizes the potential outcomes of a random variable capable of assuming one of K categories, with the likelihood of each category being explicitly defined. Categorical distributions are affiliated with a wide-ranging class of multinomial distributions [46]. Multinomial distributions by themselves are a generalization of binomial distributions. When modulating the behavior of random variables, that are limited to intervals of finite length, the beta distribution [46] is considered an appropriate distribution to model these parameters, as it is able to control the shape of the underlying distribution. For this reason, in Bayesian inference [49], the beta distribution is the conjugate prior probability distribution for the binomial distribution [46], as well as for Bernoulli and geometric distributions.

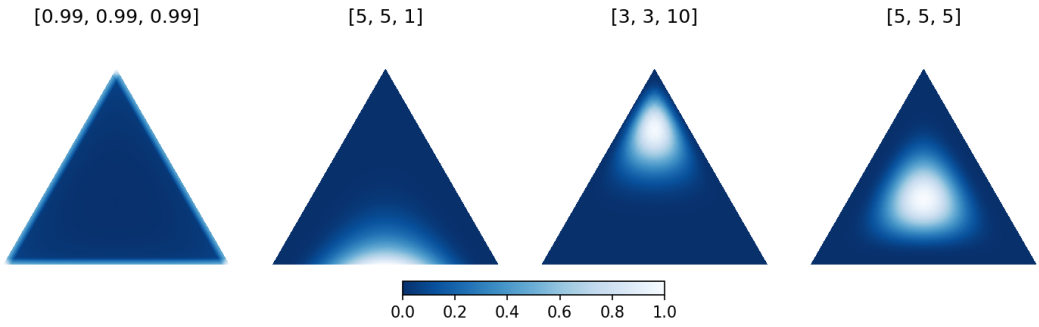


Figure 2.8: Visual representation of multiple examples of three-dimensional Dirichlet probability distributions with parameters $\boldsymbol{\alpha} = [\alpha_1, \alpha_2, \alpha_3]$, confined into a 2-Simplex. Image extracted from: gregorygundersen.com

The Dirichlet distributions are part of a family of prior distributions for the parameters of multinomial distributions [46]. In Bayesian inference, the Dirichlet distribution is essentially a generalization of the

beta distribution for probability distributions of the multinomial family [48]. For instance, the parameters of a K-dimensional categorical distribution, $\boldsymbol{p} = (p_1, p_2, \ldots, p_K)$, can be distributed over a $(K-1)$-Simplex, as exemplified in Figure 2.8, due to the related constraints, $0 \leq p_k \leq 1, \forall k \in \{1, \ldots, K\}$ and $\sum_{k=1}^{K} p_k = 1$. Consider that $\boldsymbol{\alpha} = (\alpha_1, \alpha_2, \ldots, \alpha_K)$ are the set of parameters $(\alpha_k > 0, \forall k \in \{1, \ldots, K\})$ of the Dirichlet distribution that models the $\boldsymbol{p}$ parameters. The normalized form [46] for this distribution is

$$\operatorname{Dir}(\boldsymbol{p}|\boldsymbol{\alpha}) = \frac{1}{B(\boldsymbol{\alpha})} \prod_{k=1}^{K} p_k^{\alpha_k - 1}, \tag{2.9}$$

defined in terms of the multivariate beta function [46], $B(\boldsymbol{\alpha})$, that can be formulated by the expression

$$B(\boldsymbol{\alpha}) = \frac{\prod_{k=1}^{K} \Gamma(\alpha_k)}{\Gamma\left(\sum_{k=1}^{K} \alpha_k\right)}. \tag{2.10}$$

The multivariate beta function (2.10) serves as a normalizing constant for (2.9), which is the Dirichlet Probability Density Function (PDF), that is formulated in terms of the so-called Euler Gamma function

$$\Gamma(z) = \int_0^{\infty} t^{z-1} e^{-t} dt, \tag{2.11}$$

which is essentially a generalization of the factorial function to non-integer numbers [46]. As previously stated, mathematically, the Dirichlet distribution is the prior distribution for multinomial distributions, such as the categorical distribution. Taking advantage of this property, modern classifier calibration techniques [47] consider the estimation of Dirichlet distributions [48] to fine-tune the outputted categorical scores, inside a Bayesian framework. Furthermore, Dirichlet distributions can be used to model the uncertainty associated with the parameters of a categorical distribution [50], if treated as random variables.

2.4.2 Estimation of a Dirichlet distribution

The *Dirichlet Fast Fit* algorithm, proposed by Minka [51], considers the maximization of the log-likelihood function that concerns a set containing training data. Essentially, the $\boldsymbol{\alpha}$ parameters of a Dirichlet distribution (2.9) can be estimated from a training set of multinomial data, $\mathcal{D} = \{\boldsymbol{p}_1, \ldots, \boldsymbol{p}_N\}$, where N represents the total amount of data points, i.e. the multiple $\boldsymbol{p}$ multinomial score vectors, used for the training procedure. Taking advantage of (2.9) and (2.10) the log-likelihood [45] is defined of the dataset

$$\log p(\mathcal{D}|\boldsymbol{\alpha}) = \log \prod_{i=1}^{N} \operatorname{Dir}(\boldsymbol{p}_i|\boldsymbol{\alpha}) = N \left(\log \Gamma \left(\sum_{k=1}^{K} \alpha_k \right) - \sum_{k=1}^{K} \log \Gamma(\alpha_k) \right) + \sum_{k=1}^{K} \sum_{i=1}^{N} (\alpha_k - 1) \log p_{i,k} \tag{2.12}$$

where $p_{i,k}$ represents the k-th parameters of the i-th multinomial vector. Given that there is no known closed-form solution to maximize (2.12), the maximum must be established through an iterative method.

In Appendix A are presented in detail all the steps of Minka's algorithm [51], which considers an alternate optimization between the precision, $\nu = \sum_{k=1}^{K} \alpha_k$, and the mean values, $\boldsymbol{m} = \left(\frac{\alpha_1}{\nu}, \ldots, \frac{\alpha_K}{\nu}\right)$, by fixing one parameter and only optimizing the other, during each iteration, in order to simplify and speed up the training process. This algorithm suppresses the need for a closed-form Dirichlet estimation.

2.4.3 Uncertainty Quantification

It is important to quantify uncertainty in object detection models as it allows one to better understand the reliability of the predictions [41]. Quantifying uncertainty can also be useful to detect and handle errors or rare cases with which the model was not trained to deal. For instance, if a model is used to detect objects in images and is highly uncertain about its predictions for a particular image, it may be an indication that there is something unusual about that image that the model was not prepared to handle.

There are two main sources of uncertainty to be considered in machine learning models [52]: data and knowledge. Data-induced uncertainty can be related to class overlap or the presence of noise in the data, while knowledge uncertainty can arise from mismatches between test and training data. According to this distinction, two different types of uncertainty [53] can be defined: epistemic and aleatoric.

Epistemic or model uncertainty is the uncertainty in the model parameters, usually as a result of the confusion about which model generated the training data [54], and can be explained with enough representative training data points. Aleatoric or observation uncertainty results from the stochastic nature of the observed input and persists in the network output despite expanded training on additional data [54].

Given a score vector representing a categorical probability $\boldsymbol{p} = [p_1, p_2, \ldots, p_K]^T$, where $\sum_{i=1}^{K} p_i = 1$ and $0 \leq p_i \leq 1\,,\ \forall i \in \{1, \ldots, K\}$, the associated classification entropy [43, 46], which can be defined as

$$\mathcal{H}(\boldsymbol{p}) = -\sum_{i=1}^{K} p_i \log p_i\,, \tag{2.13}$$

quantifies the amount of confusion represented by that vector $\mathbf{p}$, measuring the underlying uncertainty. Therefore, the entropy (2.13) is a metric that allows for the quantification of the uncertainty associated with a discrete probability distribution. The Kullback-Leibler (KL) divergence [46] is considered a common approach to measure the difference between two continuous distributions, which is expressed as

$$\mathcal{D}_{KL}\left(P \,||\, Q\right) = \int_{-\infty}^{+\infty} p(x) \log\left(\frac{p(x)}{q(x)}\right) dx\,, \tag{2.14}$$

where P and Q represent two continuous probability distributions, with p and q as the respective PDFs. One known technique [3] used to quantify uncertainty, involving the KL divergence (2.14), is to define Q as baseline distribution that represents a state of maximum uncertainty, measuring its distance to the probability distribution P, which can represent the distribution of confidence across a certain domain.

2.5 Active Visual Perception

Active perception is a type of active learning [4] where an agent gathers information from sensors and combines the current state of information with prior knowledge of the world to determine the next best action. This approach can be used with various types of sensors and stimuli (e.g. tactile, auditory, and olfactory), yet this study focuses on using visual sensory information to perform visual-cognitive tasks.

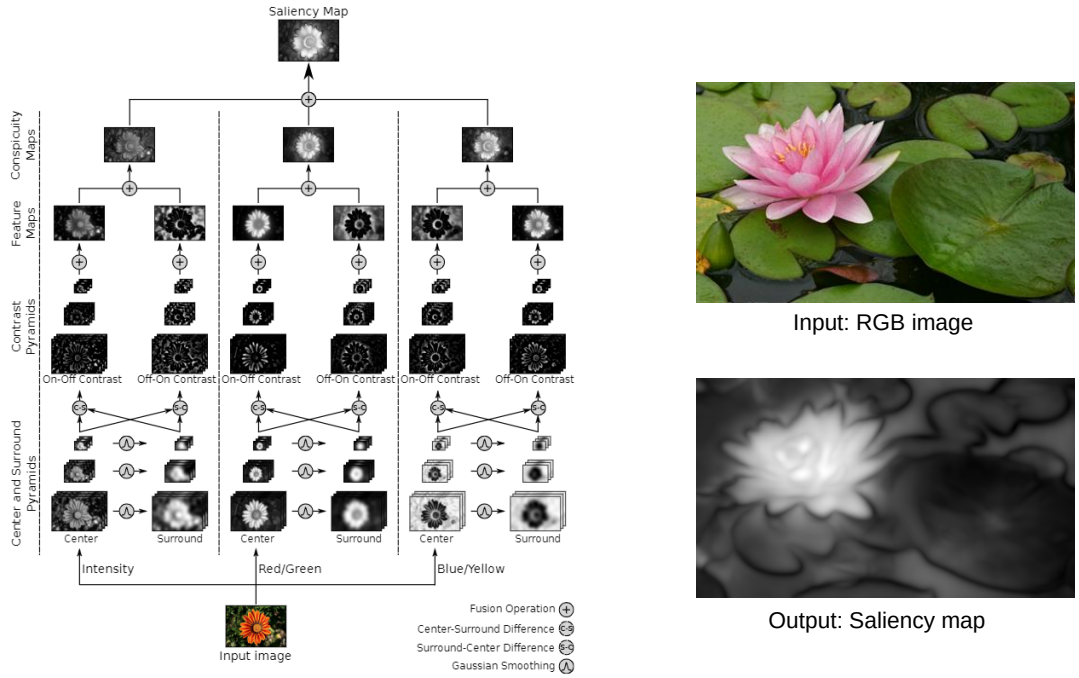


Figure 2.9: Overview of the bottom-up saliency system VOCUS2 [15], inspired by Itti's iLab Neuromorphic Vision Toolkit (iNVT) [9]. An input image is divided into multiple color and intensity contrast channels. These channels lead to center-surround pyramids, obtained through Gaussian smoothing. Then, each pyramid's center-surround differences are computed, at a multi-scale level, to generate on-off and off-on contrast pyramids. The new multi-scale contrast pyramids are then fused on feature maps, from which are generated channel-wise conspicuity maps. Finally, the conspicuity maps of the diverse color contrast channels are fused on a single master map, which is commonly known as a saliency map.

When considering visual tasks, the next move consists of a gaze shift toward a different focal point [19]. Therefore, in this context, the active decision consists of selecting the most promising region, to where the gaze should then be shifted [55, 56]. To guide the selection process Figueiredo et al. [57] considered the usage of acquisition functions, such as the probability of improvement or the expected improvement, to choose the point that maximizes a task-specific function [3], through a stochastic optimization process. This can be accomplished through the application of appropriately pre-defined metrics, such as the entropy (2.13), the KL divergence (2.14), or the probability of a particular class.

Recently developed models, that consider either uniform Cartesian [28,29] or foveal [22,32,34] visual systems, make use of CNNs to extract and process bottom-up, top-down, or both types of attentional

features, which can be used to spot regions of interest over the entire visual field. Alternative approaches [3, 25] exploit the rich semantic information outputted by modern deep object detection models, building context grids that map the information collected from multiple object predictions, distributed around the whole field of view. Some other approaches, e.g. [25, 58], even consider modeling the full extent of the humanoid body behavior, allowing not only eye movements but also full head and neck mobility.

2.5.1 Saliency-Based Framework

The traditional Itti-Koch attention model [9], also known as the iNVT, inspired multiple modern approaches to human visual cognition, and among those is VOCUS2 [15], an improved and modernized version of this classical model. This model comprises multiple necessary adaptations that maximize the performance of the model, with special emphasis on the scale-space, enabling a flexible center-surround ratio definition. VOCUS2 [15] is not only coherent in structure but also fast and, as a typical bottom-up model, performs saliency computations solely at the pixel level.

A large portion of saliency systems perform computations based mainly on intensity and color features. This corroborates with background studies on human perception since color is one of the basic features that guide visual attention [59]. Opponent theory of human perception states that there are three opponent channels in the human visual system: red versus green, blue versus yellow, and black versus white. In summary, VOCUS2 [15] starts by converting the input RGB image into an opponent-color space with channels for intensity, red-green, and blue-yellow contrasts, which can be defined as

$$I = \frac{R+G+B}{3}\,, \quad RG = R-G\,, \quad BY = B - \frac{R+G}{2}\,. \tag{2.15}$$

Contrary to its predecessor, VOCUS [37], which had a bottom-up and a top-down part, VOCUS2 [15] is based on a pure bottom-up approach. In each channel, two image pyramids (a center pyramid and a surround pyramid) are calculated, from which the on-off and off-on conspicuity contrasts are computed, as schematized in Figure 2.9. A saliency map is then generated through the fusion of scale pyramids from the feature channels. This overall structure is in correspondence with the FIT [36]. From the generated saliency map, the high contrast regions can be interpreted as RoIs and be defined as promising focal points. It is not very clear whether the quality of the saliency maps is guaranteed when foveated images are used as input to the model. One objective of this work is precisely to understand whether a method in correspondence with FIT, such as VOCUS2 [15], can be applied in a foveal context.

Zero-shot learning is a type of machine learning in which a model can recognize and classify objects it has never seen before, based on a description of the object's characteristics [29]. This is achieved by training the model on a large dataset of labeled images and providing it with a set of attributes that describe the characteristics of the objects in the dataset. The model can then use these attributes to

classify new unseen objects based on their characteristics, without the need for additional training data. The IVSN [29] is a zero-shot model, that takes advantage of the immense power of the CNNs to extract top-down high-level salient features from a visual scene, in order to perform a visual target search [17]. Similarly to the traditional Itti-Koch attention mechanism [9], the proponents of IVSN consider both WTA and IOR mechanisms when iteratively selecting the most promising regions from the field of view.

2.5.2 Semantic-Based Framework

Modern active visual perception approaches [3, 25, 60] consider the incorporation of the rich semantic information extracted from the visual scene by modern semantic object detectors [18]. Semantic-based approaches generally attempt to map the semantic content of the surrounding environment by associating and fusing classifier scores into their respective localization on an appropriate spatial representation. For instance, in the work of Druon et al. [25] the semantic context from the objects present in the scene is learned by transforming visual information into an intermediate representation called context grid [57] which essentially represents how much an object at a given location is semantically similar to a target object. Moreover, Dias et al. [3] consider sequentially fusing semantic information, extracted across multiple saccades, to perform scene exploration with foveal vision and active semantic perception [4].

3

Methodology

Contents

The central objective of this work is to establish how well semantic-based active perception, adopted by Dias et al. [3] in their model, is able to complete visual tasks that are regularly performed (often even unconsciously [5,61]) by human beings, namely scene exploration and visual target search, and to compare its accuracy with a biologically inspired model [15] (derived from past studies and developments in both the fields of neurology and psychology [9]) that attempts to imitate the human cognitive system.

This chapter aims to fully explain the methodological apparatus, detailing underlying mathematical models and mechanisms that govern the different blocks that compose the artificial perceptive model.

3.1 General Approach

The underlying principle that sustains the methodological approach to be presented (fully schematized in Fig. 3.1) is that, when performing a vast majority of human activities, particularly visual tasks, there is a great advantage in using all past sensory information [4] to plan the next move. Such a strategy can efficiently minimize the number of actions (i.e. gaze shifts) necessary to fully complete a given task. Following the approach proposed by Dias et al. [3] in this work is considered the usage of the rich semantic information, outputted by a state-of-the-art deep object detection model (e.g. YOLOv3 [1]), to iteratively direct the gaze toward regions of interest when performing humanoid visual tasks, namely scene exploration and visual search, while using artificial foveal vision [8] to emulate the visual field.

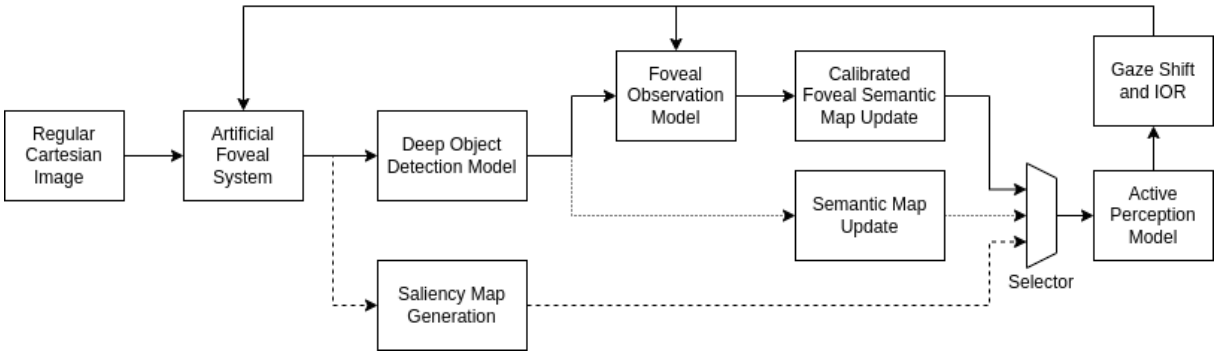


Figure 3.1: Block diagram that explains the general methodological approach to the visual tasks. An image is foveated in some initial point to simulate the human visual field. The foveal image is fed to an object detection model that may generate multiple bounding-boxes paired with classification scores, which are then used to update the semantic information maps [3], either with or without foveal calibration. Adding to these two maps, a saliency map of the scene is also generated and supplied to the active perception block that, depending on the selected method, considers one of the available maps to predict the next best focal point. The image is then foveated in the selected new focal point, simulating a saccade, IOR is applied in the different maps, and the process is repeated until a certain terminal condition is met.

This methodology assumes the usage of an artificial foveal system [8,32,34] to foveate regular Cartesian images. As described in Section 2.1.3, Almeida's model [8] is based on a combination of Gaussian and Laplacian pyramids, taking advantage of space-variant low-pass spectral filters to introduce differ-

ent levels of blur on the distinct regions of the image. One of the advantages of this foveation model is that it is fast, which makes it applicable in a real-time context. In the context of robotic vision, the are multiple state-of-the-art structures [57] that can be used to represent the spatial environment. For the sake of simplicity, a fixed field of view is considered, meaning that the agent can only perform simple ocular movements without dynamically changing the configuration and the boundaries of the visual field. Taking this limitation into account, a typical two-dimensional Cartesian representation (namely, an occupancy grid [57]) is applied. In semantic exploration, the goal is set on accumulating and mapping the information that is sequentially provided by a state-of-the-art object detection model [18] in the form of object predictions outputted at each fixation. Consider now a semantic map that aggregates categorical information scores. These scores represent the accumulated knowledge about the semantic content distributed across the visual field, until the timestamp t, comprised on a map of categorical distributions

$$\mathbf{M}_t(x,y) = \mathbf{p}_t^{x,y} = \left[p_{t,0}^{x,y}, p_{t,1}^{x,y}, \ldots, p_{t,K}^{x,y}\right]^T \in \Delta^K \,, \tag{3.1}$$

where $p_{t,k}^{x,y} = P\left(C^{x,y} = k | \mathcal{I}_{1:t}, x_{1:t}, y_{1:t}\right)$ represents the posterior probability of the object located in coordinates (x,y) being an instance of class $k \in \mathcal{C}$, assuming that $\mathcal{C} \subseteq [0,1,\ldots,K]$ represents the total set of possible classes. The posterior probabilities are conditioned by the semantic information contained in all sets of object predictions $\mathcal{I}_{1:t}$, extracted at the respective $(x_{1:t}, y_{1:t})$ focal coordinates. As suggested in the diagram presented in Figure 3.1, this map can be updated using either the calibrated (S') or raw non-calibrated (S) categorical scores, that are associated with proximal detections, trough an appropriate classifier fusion method [47,62], to be presented in Section 3.3. The calibration of categorical scores is attained through a mechanism, referred to as the foveal observation model by Dias et al. [3], which is further detailed in Section 3.2. Following the flow of information in the diagram presented in Figure 3.1, new calibrated scores can then be sequentially fused in a map that represents the categorical distributions (3.1) which express the semantic content of the scene, as learned by the model. This map is appropriately named as the Modified Kaplan map by Dias et al. [3]. An alternative approach consists of simply circumventing the foveal observation model and directly updating a semantic map (3.1) with the scores that are outputted by the object detection model. Following the nomenclature imprinted by Dias et al. [3], from now on the semantic map that is updated with raw classifier scores will be referred to as the Kaplan map, honoring the author of the rule [62] that will be applied during the successive updates.

Through the incorporation of a biologically inspired saliency model, as schematically illustrated in Figure 3.1, it becomes possible to evaluate the performance of a model that mimics key features of the human cognitive system [9], by selecting the most promising RoIs on the basis of pure salient features, during the completion of the different visual tasks. In summary, as depicted in Figure 3.1, the active perception model, which will be exposed in Section 3.4, is able to select and process data from one of three different sources to predict the next focal point, taking into consideration the scene's full extension.

3.2 Score Calibration for Object Detectors

As portrayed in the diagram presented in Figure 3.1, the enunciated methodology hinges on localized semantic information, which is extracted from the scene by an object detector. Moreover, this information should, in principle, also be calibrated [3] to account for the uncertainty imposed by the foveal sensor.

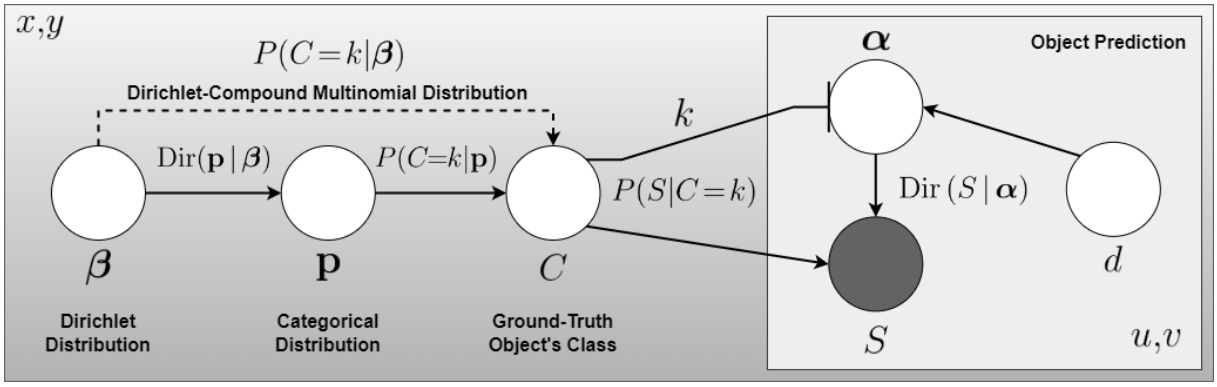


Figure 3.2: Representation of the dependencies between variables that participate in the methodology [3], in the form of a Bayesian network using the plate notation, explaining the foveal score calibration technique.

A fine-tuned calibrated K-class probabilistic classifier, akin to those incorporated within contemporary deep object detection models [18], correctly quantifies the level of uncertainty or confidence inherent to its instance-wise predictions [47]. For this reason, Dias et al. [3] apply a known technique [48], considering the scores (S), outputted by the classifier, as random variables that can be realized by a Dirichlet distribution $S \sim \mathrm{Dir}(\boldsymbol{\alpha})$, as suggested in Figure 3.2. Moreover, the incorporation of a foveal visual system generates extra data-induced uncertainty [53] on the outputs of the object detection model. This assumption comes from the fact that most classifiers are solely trained with datasets containing regular Cartesian images, not modeling the geometrical transformations [31] that characterize the foveal image inputs. These transformations are related to the space-variant distortion that accompanies the focal distance, represented as d in Figure 3.2, in relation to some detection observed at a given instant.

Taking this approach into consideration, Dias et al. [3] proposed a foveal calibration mechanism that aims at capturing the variability of S, for each class $k \in \mathcal{C}$, based on the geometrical features of the new foveated inputs. To infer the distribution $p\,(C = k|S, d)$, where C represents the actual object's class, the distribution $p\,(S|C = k, d)$ is learned and then applied in a simple Bayes classifier [52], expressed as

$$p(C = k|S, d) = \frac{p(S|C = k, d)\, p(C = k|d)}{p(S|d)} \propto \mathrm{Dir}\,(S|\boldsymbol{\alpha}_{k,d})\ , \tag{3.2}$$

permitting the determination of S', which are the desired calibrated scores. Essentially, by taking advantage of the Bayes theorem [49] it becomes possible to determine the posterior probability of each class k, given the classifier's evidence S, by estimating the likelihood of S, assuming that indeed $C = k$,

all under prior knowledge of the focal distance. It is assumed a uniform prior distribution $p(C = k|d)$, considering that an object can be arbitrarily located within the confinements of the visual field and, for this reason, there is no direct correlation between the focal distance d and the actual class of the object.

3.2.1 Semantic-Based Object Detection

In order to generate predictions, as depicted in Figure 3.1, it must be selected a semantic-based object detection model. At a given time stamp t, with the gaze fixed upon a certain region of the visual field, suppose that a selected semantic object detector is able to output a set of object predictions $\mathcal{I}_t$, containing in total L_t detections. Each detection $I_{t,l} \in \mathcal{I}_t$, where $l \in \{1, \ldots, L_t\}$, represented by a plate in Figure 3.2, consists of a bounding-box $\mathcal{B}_{t,l}$ (defined by the coordinates of the top-left and bottom-right corner pixels) and a normalized vector containing categorical probability scores, which is formalized as

$$S_{t,l} = (s_{t,l,1}, s_{t,l,2}, \ldots, s_{t,l,K}) \ , \tag{3.3}$$

where $0 \leq s_{t,l,k} \leq 1 \, , \forall k \in \{1, \ldots, K\}$ and $\sum_{k=1}^{K} s_{t,l,k} = 1$. Every single score from (3.3) can be interpreted as the posterior probability of the corresponding class, inside the region that is delimited by the respective bounding-box $\mathcal{B}_{t,l}$, which can then be directly fused on a semantic map, as in Figure 3.1.

3.2.2 Foveal Observation Model

The foveal observation model hinges on a critical understanding of the distortion's spatial distribution within the visual field. This distribution exhibits a discernible radial pattern, characterized by a progressive increase in blur level as one moves outward from the central foveal region towards the peripheral region. For this reason, a score vector $S_{t,l}$ (3.3) is expected to present less entropy (3.20) when the detection's bounding-box is located near the center of the fovea [3], typically increasing to higher values as its location drifts deeper into the peripheral regions. Consequently, the foveal correction mechanism is strategically devised to leverage this spatial relationship between objects and the focal point, effectively accounting for the underlying epistemic uncertainty associated with the object detection model. A multinomial score vector $S_{t,l}$ can be associated with a Dirichlet prior that depends on the location of the detection with respect to the center of the foveal region, which can be measured in the form of a distance, $d_{t,l}$, between the center of the respective bounding-box and the focal point.

Essentially, the foveal observation model consists of multiple sets of Dirichlet distribution parameters, specifically a structure that contains $K \times N$ arrays of alpha parameters. This means that there is a set of parameters for each class in all N distance levels. These levels are not more than mere discretizations of the distance measurement $d_{t,l}$ that grant the existence of a finite number of compartmentalized Dirichlet distributions that are trained to model the $p\left(S_{t,l}|C = k, d_{t,l}\right)$ probability distribution, as in (3.2).

Given the unavailability of a closed-form maximum-likelihood estimation for Dirichlet distributions, each set of alpha parameters is estimated by fitting multiple detection data, consisting of multiple score vectors (3.3) (outputted by the detector when operating over a large dataset, containing multiple foveal images) that are sorted by distance level. This is achieved with an efficient iterative method [51] that has been already presented in Section 2.4.2. The process culminates in multiple sets of parameters

$$\boldsymbol{\alpha}_{k,d_{t,l}} = \left(\alpha_{k,d_{t,l},1}, \alpha_{k,d_{t,l},2}, \ldots, \alpha_{k,d_{t,l},K}\right) \tag{3.4}$$

that, in a sense, are able to capture both aleatoric and epistemic uncertainties [48] derived from the distortion, characteristic of the (artificially imposed) foveal nature of the visual field, that provokes confusion on the output of the object detection model, which has been pre-trained on regular Cartesian images.

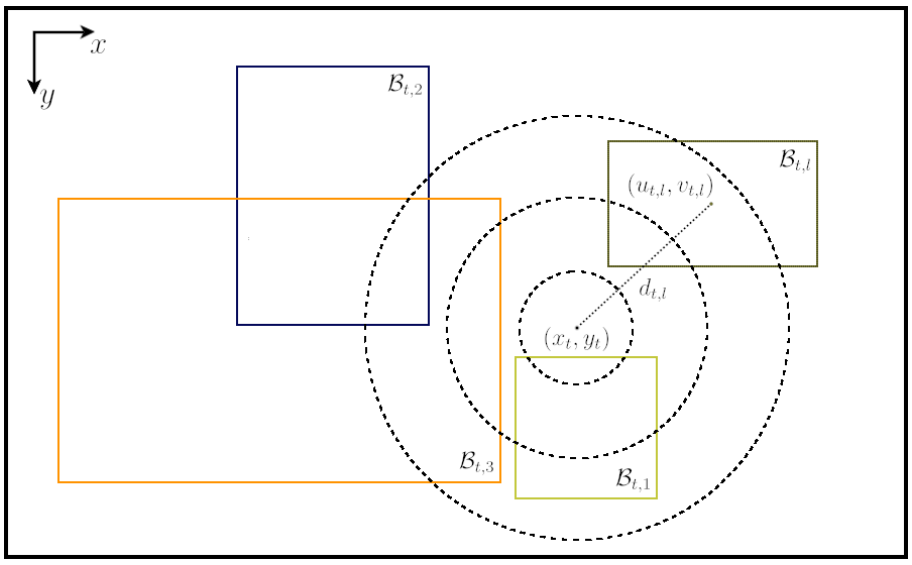


Figure 3.3: Illustration of a foveal scene (fixation) where YOLOv3 was applied, outputting L_t object predictions. The focal distance $d_{t,l}$, between the center of a bounding-box $\mathcal{B}_{t,l}$ and the focal point, is also represented.

Considering that the fovea may not be equally distributed on both horizontal and vertical dimensions, the distance between the center of the fovea (x_t, y_t) and the center of the bounding-box $\mathcal{B}_{t,l}$ that delimits a detection $I_{t,l}$ is determined using the Mahalanobis distance in order to generalize the approach [63], hypothetically allowing for an elliptically shaped foveal region. In the foveal context, the Mahalanobis distance [64] depends directly on the relative coordinates $(u_{t,l}, v_{t,l})$ of a detection $I_{t,l}$ to the focal point, in accordance with the illustration presented in Figure 3.3, being defined by the expression

$$d_{t,l}(u_{t,l}, v_{t,l}) = \sqrt{[u_{t,l}, v_{t,l}]\Sigma^{-1}[u_{t,l}, v_{t,l}]^T}\,, \quad \text{where} \quad \Sigma = \begin{bmatrix} \sigma_x & 0 \\ 0 & \sigma_y \end{bmatrix} \tag{3.5}$$

is known as the covariance (or weight) matrix, a diagonal matrix characterized by the weights σ_x and σ_y that define how extendable the foveal region is across both the horizontal and vertical dimensions,

respectively. Assuming that for a certain object detection $I_{t,l}$ the pixel coordinates of the center of the bounding-box $\mathcal{B}_{t,l}$ are (x, y), the local coordinates in respect to the center of the fovea can be established as $(u_{t,l}, v_{t,l}) = (x - x_t, y - y_t)$. With a maximum distance that properly fits the dimensions of the images at hand, it becomes possible to organize the detections into N distance levels, defining proper superior and inferior limits, to then find in which of the intervals the measured distance $d_{t,l}$ falls in. In order to properly counter the uncertainty imposed by different levels of distortion, a new set of classification scores $S'_{t,l}$, calibrated to more accurately represent the scores associated with an object prediction, given its focal distance $d_{t,l}$ (3.5), that expresses the likelihoods of $S_{t,l}$ (3.3) for each class $k \in \mathcal{C}$, is represented as

$$S'_{t,l} = \left(s'_{t,l,1}, \ldots, s'_{t,l,K}\right) = \frac{1}{D}\left[p\left(S_{t,l}|C = 1, d_{t,l}\right), \ldots, p\left(S_{t,l}|C = K, d_{t,l}\right)\right]^T, \tag{3.6}$$

where $D = \sum_{k=1}^{K} p\left(S_{t,l}|C = k, d_{t,l}\right)$ is the normalization factor which grants that $\sum_{k=1}^{K} s'_{t,l,k} = 1$. Taking advantage of the pre-trained parameters (3.4), the probability distribution $p\left(S_{t,l}|C = k, d_{t,l}\right)$ can be modeled as a realization of a Dirichlet through the Bayes classifier (3.2), resulting in the new formulation

$$S'_{t,l} = \frac{1}{D}\left[\mathrm{Dir}\left(S_{t,l}|\boldsymbol{\alpha}_{1,d_{t,l}}\right), \mathrm{Dir}\left(S_{t,l}|\boldsymbol{\alpha}_{2,d_{t,l}}\right), \ldots, \mathrm{Dir}\left(S_{t,l}|\boldsymbol{\alpha}_{K,d_{t,l}}\right)\right]^T, \tag{3.7}$$

where, once again, $D = \sum_{k=1}^{K} \mathrm{Dir}\left(S_{t,l}|\boldsymbol{\alpha}_{k,d_{t,l}}\right)$ is the normalization factor. The Dirichlet likelihoods can be estimated with (2.9), substituting the probability vector $\mathbf{p}$ with $S_{t,l}$, which are the scores directly outputted by the object detection model (YOLOv3) and being calibrated in the foveal correction model.

3.3 Fusion Model for Semantic Information Maps

In a Bayesian context [41], the parameters $\mathbf{p}$ themselves can be treated as random variables with associated uncertainty [46], and therefore, selecting a prior distribution that captures the initial beliefs about each parameter of the categorical distribution it is possible to encode attach the underlying uncertainty. One simple and convenient solution is to model the parameters $\mathbf{p}$ as realizations of a Dirichlet distribution, $\mathbf{p} \sim \mathrm{Dir}(\boldsymbol{\beta})$, with parameters $\boldsymbol{\beta} = (\beta_0, \beta_1, \ldots, \beta_K)$, as entailed in Figure 3.2. Mathematically, the Dirichlet distribution is the conjugate prior for the categorical distribution, meaning that when a Dirichlet prior is combined with a categorical likelihood, it results in a posterior distribution that is also a Dirichlet.

Taking advantage of these properties, in conformity with the Bayesian network presented in Figure 3.2, a new map containing the parameter's estimates at time t on (x_m, y_m) coordinates is defined as

$$\mathbf{B}_t(x_m, y_m) = \boldsymbol{\beta}_t^{x_m, y_m} = \left[\beta_{t,0}^{x_m, y_m}, \beta_{t,1}^{x_m, y_m}, \ldots, \beta_{t,K}^{x_m, y_m}\right]^T, \tag{3.8}$$

being able to capture the uncertainty of the class posterior scores $p_{t,k}^{x_m, y_m}$, arranged on the semantic map (3.1), through what is called in Bayesian statistics as the second moment. The variance (i.e. second

moment) measures the dispersion associated with the categorical parameters [52]. Notice that in (3.8) the (x, y) coordinates have been replaced by (x_m, y_m). This substitution comes from the fact that, in order to facilitate the localized classifier fusion process, the visual field is divided into multiple uniform regional divisions, in the form of an $X \times Y$ grid of rectangular cells. Each cell on this grid is defined and referred to with a respective pair of (x_m, y_m) coordinates. Following the previous definition, the existence of (3.8) allows for inference of the posterior categorical probability [3, 46] on (x_m, y_m) coordinates as

$$\mathrm{M}_{t,k}(x_m, y_m) = p_{t,k}^{x_m,y_m} = P\left(C = k \,|\, \boldsymbol{\beta}_t^{x_m,y_m}\right) = \frac{n!}{\left(\sum_{i=0}^{K} \beta_{t,i}^{x,y}\right)^n} \prod_{i=1}^{K} \left\{ \frac{\left(\beta_{t,i}^{x,y}\right)^{n_i}}{n_i!} \right\} = \frac{\beta_{t,k}^{x,y}}{\sum_{i=0}^{K} \beta_{t,i}^{x,y}} \tag{3.9}$$

given that $P\left(C = k \,|\, \boldsymbol{\beta}_t^{x_m,y_m}\right)$ can be realized by a Dirichlet-compound multinomial distribution [50], where $n = \sum_{i=0}^{K} n_i$, thereby assuming $n_k = 1$ $(k \in \mathcal{C})$ and $n_i = 0, \forall i \in \mathcal{C} \setminus \{k\}$, as conveyed by Figure 3.2. This shortcut allows one to overlook the linked dependencies between the Dirichlet distribution parameters (3.8), the categorical parameters (3.1), and the actual class of the observed object, allowing for a direct association between semantic information (3.8) and the actual class of the objects.

For the sake of simplicity, consider now that some arbitrary cell (x_m, y_m) has been fixed and that $\beta_{t,k}$ (3.8) are the estimated parameters of the underlying Dirichlet distribution, at some instant t, that models the uncertainty attached to the categorical parameters $\mathbf{p}$ on that specific cell. There are multiple update rules (e.g. product rule, sum rule [65]) that can be used when sequentially fusing the scores outputted by a classifier. Following the work developed by Kaplan et al. [62], that compares Bayesian classifier fusion approaches with a proposed rule, to which Dias et al. [3] conveniently refers to as Kaplan's update rule, formulated for a fixed (x_m, y_m) cell, using the scores outputted by the detector for a given $I_{t,l}$, as

$$\beta_{t+1,k} = \frac{\beta_{t,k} \left(1 + \dfrac{s_{t,l,k}}{\sum_{j=0}^{K} \beta_{t,j} s_{t,l,j}} \right)}{1 + \dfrac{\min_j s_{t,l,j}}{\sum_{j=0}^{K} \beta_{t,j} s_{t,l,j}}} \,. \tag{3.10}$$

The Kaplan rule (3.10) is applied when fusing new score vectors (3.3), associated with the detections that either totally or partially overlap the (x_m, y_m) cell, as it is able to return a posterior Dirichlet distribution, through a moment-matching approach [62], that fits the actual posterior distribution of $\mathbf{p}$ parameters, properly modeling the underlying uncertainty. Suppose multiple detections overlap in the (x_m, y_m) coordinates. In the work of Simões [63], there are considered multiple update normalization techniques (e.g. most probable box, artificial increments). Nonetheless, it was shown that, without this normalization, the scene exploration results can actually be decisively improved. The most plausible explanation for this observation is that, since not even a single detection is ignored, there is no loss of information. Therefore, in this work, in a multiple overlapping detections situation, the Kaplan rule (3.10) is sequentially applied, in no particular order, with every score vector (3.3) coming from those detections.

3.3.1 Integration of the Background Class

Following the methodology applied in previous works [3, 63], given the fact that a cell may not be overlapped with any object available in the scene, a vector (3.3) should also aggregate the confidence scores that represent the actual probability of not containing objects inside the predicted bounding-box.

For this reason, the background class is integrated into the model by defining its respective score $s_{t,l,0}$ as the value that the classifier would output in a situation of maximum uncertainty, which corresponds to a uniform distribution $\left(\frac{1}{K+1}\right)$, redefining the categorical scores (3.3) generated by the object detector

$$s_{t,l,k} = \begin{cases} \frac{K}{K+1}\, s_{t,l,k}\,, & k = \{1, \dots, K\} \\ \frac{1}{K+1}\,, & k = 0 \end{cases} \tag{3.11}$$

where the normalization factor $\frac{K}{K+1}$ ensures that $\sum_{k=0}^{K} s_{t,l,k}$ still remains a unit-sum after the inclusion of the background class. Despite the intention, with the proposed approach (3.11) no real information about the model's confidence in the existence of an object (or the lack thereof) is attached to the semantic map. It is possible that precious information provided by object detectors [18], in the form of objectness scores, is being neglected by the proponents of the semantic model [3] when defining an inclusive methodology that also contemplates the confidence in the existence of objects inside each cell's confinements.

3.3.2 Incorporation of Foveal Calibrated Scores

Kaplan's update rule [62] can also accommodate the foveal calibrated scores (3.6), as it is possible to directly replace (3.3) with the corrected foveal scores (3.6), resulting in a new slightly different formulation

$$\beta_{t+1,k} = \frac{\beta_{t,k}\left(1 + \frac{s'_{t,l,k}}{\sum_{j=0}^{K} \beta_{t,j} s'_{t,l,j}}\right)}{1 + \frac{\min_j s'_{t,l,j}}{\sum_{j=0}^{K} \beta_{t,j} s'_{t,l,j}}}\,, \tag{3.12}$$

attained after properly extending (3.6) to aggregate the background class, using the (3.11) rule. This formulation consists of what Dias et al. [3] refer to as the Modified Kaplan rule, which is the rule used when specifically updating the Modified Kaplan map. Each individual likelihood from (3.6) is sampled as

$$s'_{t,l,k} = \mathrm{Dir}\left(S_{t,l} | \boldsymbol{\alpha}_{k,d_{t,l}}\right) = \frac{1}{B(\boldsymbol{\alpha}_{k,d_{t,l}})} \prod_{j=1}^{K} s_{t,l,j}^{\alpha_{k,d_{t,l},j}-1}\,, \tag{3.13}$$

on the foveal observation model, in conformity with (3.7), from the PDF of the $\mathrm{Dir}\left(S_{t,l} | \boldsymbol{\alpha}_{k,d_{t,l}}\right)$ distribution (2.9). This small adaptation is contemplated in Figure 3.1, on the straight line path that includes the foveal observation model, presented in Section 3.2.2, generating the calibrated scores (3.6) that are used to sequentially update the map that feeds semantic information to the active perception model.

3.4 Active Perception using Semantic Information Mapping

Consider now the active perception model, which operates based on the information collected on the semantic map (3.8), at a given timestamp t, to find the most promising cell (x^*_{t+1}, y^*_{t+1}) according to a pre-defined task-specific metric, in order to select the region where the gaze direction is to be shifted at the $t+1$ instant. Following the methodology proposed by Dias et al. [3], it is considered a full simulation of $\mathbf{B}_t$ (3.8) map update for each possible next focal cell (x'_m, y'_m) where the fovea could be centered next, obtaining a set of new point-of-view dependent maps that aggregate accumulated semantic information

$$\bar{\mathbf{B}}^{x'_m, y'_m}_{t+1}(x_m, y_m) = \mathbb{E}\left[\mathbf{B}_{t+1}(x_m, y_m) \,|\, \mathbf{B}_t(x_m, y_m), x'_m, y'_m\right], \tag{3.14}$$

in order to predict the possible states of knowledge presented by the semantic map. From (3.14), it is possible to infer metric estimations, since each $\bar{\mathbf{B}}^{x'_m, y'_m}_{t+1}$ comprises the estimated Dirichlet distributions, after the updates, assuming that the gaze was shifted to the center of (x'_m, y'_m). This means that the active perception mechanism has to select one out of $X \times Y$ possible scenarios, based on the knowledge that results from each simulation. To solve (3.14), it is considered the computation of the expected value for the scores that should be outputted by the object detector for the (x_m, y_m) cell coordinates if the next focal point was set upon the center of a (x'_m, y'_m) cell, through the expansion of the posterior distribution

$$\begin{aligned} p\left(S^{x_m, y_m} \,|\, x'_m, y'_m, \boldsymbol{\beta}_t^{x_m, y_m}\right) &= \sum_{k=0}^{K} p\left(S^{x_m, y_m}, C = k \,|\, x'_m, y'_m, \boldsymbol{\beta}_t^{x_m, y_m}\right) \\ &= \sum_{k=0}^{K} p\left(S^{x_m, y_m} \,|\, C = k, x'_m, y'_m, \boldsymbol{\beta}_t^{x_m, y_m}\right) P\left(C = k \,|\, x'_m, y'_m, \boldsymbol{\beta}_t^{x_m, y_m}\right) \\ &= \sum_{k=0}^{K} p\left(S^{x_m, y_m} \,|\, C = k, x'_m, y'_m\right) P\left(C = k \,|\, \boldsymbol{\beta}_t^{x_m, y_m}\right), \end{aligned} \tag{3.15}$$

where the individual categorical probabilities $P(C{=}k \,|\, \boldsymbol{\beta}_t^{x_m, y_m})$ can be correlated with the knowledge, gathered in (3.8), through (3.9), as $p_{t,k}^{x_m, y_m}$. The steps presented in (3.15) take advantage of the conditional independence between S^{x_m, y_m} and $\boldsymbol{\beta}_t^{x_m, y_m}$ for a given class k. It also hinges on the fact that the class, to which the object contained in a (x_m, y_m) cell belongs, is not constrained by the location of the center of the fovea (x'_m, y'_m) at the $t+1$ timestamp. Taking advantage of the linearity of the expectation operator applied to (3.15) in order to find the expected classifier output scores, it is possible to assert

$$\bar{S}^{x_m, y_m} = \mathbb{E}\left[S^{x_m, y_m} | x'_m, y'_m, \boldsymbol{\beta}_t^{x_m, y_m}\right] = \sum_{k=0}^{K} \mathbb{E}\left[S^{x_m, y_m} | C = k, x'_m, y'_m\right] P\left(C = k \,|\, \boldsymbol{\beta}_t^{x_m, y_m}\right), \tag{3.16}$$

meaning that the computation of predictive scores is hanging on the determination of $\mathbb{E}\left[S^{x,y} | k, x', y'\right]$ in each iteration. Consider now that $d(u', v')$ represents the distance level associated with a given (x_m, y_m)

cell in relation to the center of the (x'_m, y'_m) cell, measured through the expression (3.5) with the relative coordinates $(u', v') = (x - x', y - y')$, where (x, y) and (x', y') are the center pixel coordinates of the referred cells, respectively. Finally, taking advantage of the relative distance $d(u', v')$ and the dynamics of the foveal observation model, the expected score vectors (3.16) are computed for each (x_m, y_m) cell

$$\bar{S}^{x_m,y_m} = \sum_{k=0}^{K} \mathbb{E}\left[\mathrm{Dir}\left(S^{x_m,y_m}|\boldsymbol{\alpha}_{k,d(u',v')}\right)\right] p_{t,k}^{x_m,y_m} = \sum_{k=0}^{K} \frac{\boldsymbol{\alpha}_{k,d(u',v')}}{\sum_{j=0}^{K} \alpha_{k,d(u',v'),j}} \frac{\beta_{t,k}^{x_m,y_m}}{\sum_{j=0}^{K} \beta_{t,j}^{x_m,y_m}} \tag{3.17}$$

which can then be applied with Kaplan's rule (3.10) to update the semantic map in order to fill (3.14) with the new estimated Dirichlet parameters. From then on, through (3.14), it is possible to determine the most promising cell, to where the gaze is to be shifted, on the basis of a given task-dependent metric.

3.4.1 Active Perception in Scene Exploration

Given that, in the context of this task, the goal is to minimize the amount of confusion, distributed all over the map, while performing a minimal number of actions (i.e. gaze shifts), the decision metrics essentially consider functions [3] that quantify the inherent uncertainty. One form of quantifying confusion consists of measuring the distance between two PDFs representing the amount of confidence associated with the presence of each class. For this purpose, the Kullback-Leibler divergence (2.14) is considered, as it is a measurement of the distance between the estimated Dirichlet distribution and a base distribution

$$F_t^{x_m,y_m} = \mathcal{D}_{KL}\left(\mathrm{Dir}\left(\boldsymbol{\beta}_t^{x_m,y_m}\right) || \,\mathrm{Dir}\left(\boldsymbol{\beta}^0\right)\right), \tag{3.18}$$

where $\boldsymbol{\beta}^0$ corresponds to the initial uniform Dirichlet distribution $(\beta_k^0 = 0.5\,, \forall k \in \{0, \ldots, K\})$, representing a state of maximum confusion, that is initially warranted to each (x_m, y_m). Intuitively, it would be desirable to select the cell that maximizes the distance between the two Dirichlet distributions [66]. Following Kurt's article [67], the Kullback-Leibler divergence [46] between the Dirichlet distribution, that represents the state of (3.8) at an instant t, and the base uniform Dirichlet distribution $\boldsymbol{\beta}^0$, which can be computed as

$$\begin{aligned}\mathcal{D}_{KL}\left(\mathrm{Dir}\left(\boldsymbol{\beta}_t^{x_m,y_m}\right) || \,\mathrm{Dir}\left(\boldsymbol{\beta}^0\right)\right) = \log\Gamma\left(\sum_{k=0}^{K}\beta_{t,k}^{x_m,y_m}\right) - \sum_{k=0}^{K}\log\Gamma\left(\beta_{t,k}^{x_m,y_m}\right) + \\ - \log\Gamma\left(\sum_{k=0}^{K}\beta_k^0\right) + \sum_{k=0}^{K}\log\Gamma\left(\beta_t^0\right) + \\ + \sum_{k=0}^{K}\left(\beta_{t,k}^{x_m,y_m} - \beta_k^0\right)\left(\psi\left(\beta_{t,k}^{x_m,y_m}\right) - \psi\left(\sum_{k=0}^{K}\beta_{t,k}^{x_m,y_m}\right)\right)\end{aligned} \tag{3.19}$$

which directly involves both the Gamma function $\Gamma(z)$, defined in (2.10), and the derived *digamma* function $\psi(z)$, also known as the *Psi* function, formulated as $\psi(z) = \frac{d}{dz}\log\Gamma(z)$. As an alternative, the

negentropy, symmetrical to the classification entropy (2.13), is an uncertainty metric that measures the total amount of confusion displayed by a categorical distribution. The negentropy can be formulated as

$$F_t^{x_m,y_m} = \sum_{k=0}^{K} p_{t,k}^{x_m,y_m} \log\left(p_{t,k}^{x_m,y_m}\right) = \sum_{k=0}^{K} \frac{\beta_{t,k}^{x_m,y_m}}{\sum_{j=0}^{K} \beta_{t,j}^{x_m,y_m}} \log\left(\frac{\beta_{t,k}^{x_m,y_m}}{\sum_{j=0}^{K} \beta_{t,j}^{x_m,y_m}}\right). \tag{3.20}$$

The higher the negentropy, the lower the cell's confusion. For this reason, both the KL divergence (3.18) and the negentropy (3.20) metrics are applied with a simple acquisition function [57] that transverses each (3.14) map, accumulating the expected results of the selected metric. With this approach, the location (x'_m, y'_m) that maximizes the sum of the expected metric results over (3.14) is selected through

$$\left(x^*_{t+1}, y^*_{t+1}\right) = \underset{x'_m, y'_m}{\operatorname{argmax}} \left\{ \sum_{y_m=1}^{Y} \sum_{x_m=1}^{X} \mathbb{E}\left[F_{t+1}^{x_m,y_m} \mid x'_m, y'_m\right] \right\}, \tag{3.21}$$

where $\mathbb{E}\left[F_{t+1}^{x,y} \mid x'_m, y'_m\right]$ represents the expected measurement of the metric F for the next iteration, i.e. at the $t+1$ timestamp, assuming that the focal point is to be relocated to the (x'_m, y'_m) cell coordinates. Finally, the last alternative metric that is proposed by Dias et al. [3], conveniently named two-peak difference, consisting of the absolute difference between the categorical scores (3.1) of the two most probable classes for each cell (x_m, y_m), inferred from the semantic map (3.8), which is expressed as

$$F_t^{x_m,y_m} = \max_k \left\{ \frac{\beta_{t,k}^{x_m,y_m}}{\sum_{j=0}^{K} \beta_{t,j}^{x_m,y_m}} \right\} - \max_{k \setminus \varkappa} \left\{ \frac{\beta_{t,k}^{x_m,y_m}}{\sum_{j=0}^{K} \beta_{t,j}^{x_m,y_m}} \right\}, \text{ where } \varkappa = \underset{k}{\operatorname{argmax}} \left\{ \frac{\beta_{t,k}^{x_m,y_m}}{\sum_{j=0}^{K} \beta_{t,j}^{x_m,y_m}} \right\} \tag{3.22}$$

represents the most probable class, out of all $\mathcal{C}$ possible classes. Following the work of Dias et al. [3], this metric is applied with a different acquisition function [57], known as the expected improvement. This function aims at maximizing the expected magnitude of the two-peaks (3.22) improvement over the map

$$\left(x^*_{t+1}, y^*_{t+1}\right) = \underset{x'_m, y'_m}{\operatorname{argmax}} \left\{ \max_{x_m, y_m} \left| \mathbb{E}\left[F_{t+1}^{x_m,y_m} \mid x'_m, y'_m\right] - F_t^{x_m,y_m} \right| \right\}, \tag{3.23}$$

which is a greedy approach, since it consists of picking the maximum absolute difference, out of all (x_m, y_m) cell coordinates, for each possible next focal point (x'_m, y'_m), selecting the most promising cell.

3.4.2 Active Perception in Visual Search

The task of visual search [17] consists of trying to direct the gaze toward a region that contains one or more objects of interest. In order to simplify the procedure and enlarge the margins of success, it is not intended for the search to be directed towards a particular object, but rather a category of objects. This means that in a scene containing multiple objects of the same target class, $C \in \mathcal{C}$, once the focus is set upon one of these objects it is accepted that the target has been found, with the task being completed.

First, it is important to take into consideration the greedy nature of visual search which, contrary to scene exploration, does not aim at identifying the most promising point of view, that may clarify the semantic content available in the visual field, but rather the most promising region, where a target might be located [36]. This means that when formulating a measure of confidence to determine the next best focal coordinates, (x^*_{t+1}, y^*_{t+1}), the priority must be given to the information self-contained in each possible next cell rather than equally distributed over all the cells of the map. It is possible to intuitively infer that, when considering moving toward a cell located at the top-left corner, the semantic information gathered in a cell located in the bottom-right corner cannot be as relevant as the information collected in the top-left corner cells, as it diminishes the importance of the spatial aspect of the decision [17]. Centering the attention of the measure around the precise location of the gaze shift forces the active perception model to attribute more relevance to the semantic information contained in the focal space. For this reason, an intuitive approach of low complexity level that considers each cell individually, consisting of identifying the cell that maximizes the posterior probability (3.1) of the class C on the whole map, has been conceived to properly guide the search algorithm toward the most promising region. Rather than minimizing some measure of uncertainty [57], e.g. (3.20), over the full map in line with the scene exploration approach [3], the active perception mechanism simply needs to transverse an updated semantic map to identify which cell presents the highest categorical score for a target class C

$$\left(x^*_{t+1}, y^*_{t+1}\right) = \underset{x_m, y_m}{\operatorname{argmax}} \left\{ p^{x_m, y_m}_{t,C} \right\} = \underset{x_m, y_m}{\operatorname{argmax}} \left\{ \frac{\beta^{x_m, y_m}_{t,C}}{\sum_{k=0}^{K} \beta^{x_m, y_m}_{t,k}} \right\}, \tag{3.24}$$

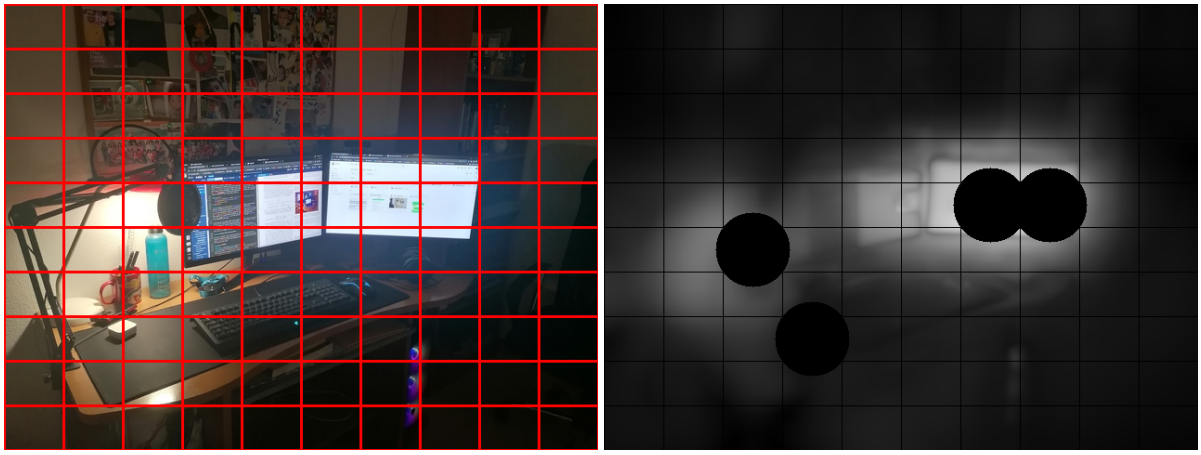
where $\beta^{x_m, y_m}_{t,C}$ is the expected parameter for the class C, extracted from a cell (x_m, y_m) contained in (3.8), at some instant t. Notice that this constitutes what can be called a non-predictive approach since it directly considers the scores comprised in either the Kaplan (3.10) or Modified Kaplan (3.12) maps, immediately after being updated with the most recent detections. To take full advantage of the perception model and properly incorporate active prediction principles, one only needs to slightly change the aspect of (3.24), obtaining a new yet similar formulation to the acquisition function (3.21), regarding (3.9), as

$$\left(x^*_{t+1}, y^*_{t+1}\right) = \underset{x_m, y_m}{\operatorname{argmax}} \left\{ \bar{p}^{x_m, y_m}_{t+1,C} \right\} = \underset{x_m, y_m}{\operatorname{argmax}} \left\{ \frac{\bar{\beta}^{x_m, y_m}_{t+1,C}}{\sum_{k=0}^{K} \bar{\beta}^{x_m, y_m}_{t+1,k}} \right\}, \tag{3.25}$$

that replaces the current Dirichlet parameter estimates $\beta^{x_m, y_m}_{t,C}$ with the expected parameters $\bar{\beta}^{x_m, y_m}_{t+1,C}$, obtained after simulating an update of the knowledge gathered about the scene at instant t with the expected classifier scores $\bar{S}^{x_m, y_m}$, computed as (3.17), using the Kaplan rule (3.10). Upon completing this procedure for all cells (x_m, y_m), the categorical probabilities can simply be inferred through (3.9) from $\bar{\mathbf{B}}^{x_m, y_m}_{t+1}$ (3.14). This means that, for each cell (x_m, y_m) of (3.14), it is considered that the next focal point is set on the same cell $(x'_m, y'_m) = (x_m, y_m)$, properly accounting for the greedy nature of the task.

3.5 Traditional Saliency Adaptation

The block diagram presented in Figure 3.1 presents one alternative path that is independent of object detectors, calibration methods, and fusion models. This path represents the semantically independent alternative approach, which considers the incorporation of a traditional biologically inspired attention model [9], which is able to provide valuable information [19] that guides the gaze selection mechanism.



(a) Grid of Cells **(b)** Saliency Map with IOR

Figure 3.4: Grid of cells (a) dividing the semantic map of Fig. 1.1, together with the saliency map (b) generated by VOCUS2 [15] from the foveal scene, after four algorithm iterations, while applying the IOR mechanism.

For this purpose, VOCUS2 [15] is incorporated in the methodological apparatus, presented in Figure 3.1 on a dashed line path. As a bottom-up saliency-based visual attention model, VOCUS2 can be considered a reformed version of Itti's traditional conspicuity model [9], generally referred to as the iNVT, that considers the extraction of low-level features, inherent to pixel-level color contrasts, in compliance with the FIT, as described in Section 2.5.1. Despite lacking top-down features, considered in modern visual attention mechanisms, such as the IVSN [29], which are essential when approximating the human neurological models to their fullness [36], VOCUS2 is still able to simulate critical aspects that guide the mechanisms governing human perception [17, 19], on the basis of pure bottom-up feature information.

Following the principles that govern the predecessor of VOCUS2 [15], namely the iNVT [9], at each iteration a WTA mechanism identifies the most salient pixel over the entire generated saliency map. Then, the active perception model applies a pixel intensity threshold, selecting only pixels with saliency-related intensity above that same threshold and ignoring the others. From then on, the remaining pixels are associated with the cell in which they are confined, in order to identify which cells represent RoIs. Finally, the number of pixels associated with each cell is counted, and then the cell (x_m, y_m) that contains the highest number of pixels is selected as the most promising region, to where the gaze is shifted next.

Each saliency map is directly generated from the visual scene that is artificially foveated around the center of the cell that is focalized in the considered iteration. Figure 3.4(b) presents a visual representation of the IOR mechanism, applied on the saliency map produced by VOCUS2 [15] after each iteration, which essentially nullifies the intensity of the pixels contained inside the previously visited (x_m, y_m) cells, from the grid displayed in Figure 3.4(a), that are identified as previously visited RoIs. The reason why the nullification of the pixel's intensity, inflicted on the saliency map, is performed in a circular shape, rather than in a rectangular shape that fits within the boundaries of each (x_m, y_m) cell, is that with that, with this technique, some small extension of an object presented in that same cell is al overlooked. Therefore, with this approach, the idea is to highlight the saliency information displayed in the center of the cells, thereby ignoring marginal information displayed near the boundaries of previously visited cells.

3.6 Task Success Evaluation

To conclude Chapter 3, this section assesses the evaluation metrics that have been considered suitable to evaluate the performances of the distinct models. The performances are measured across the iterations, as the gaze direction is shifted across the field of view during the completion of the tasks [17, 26].

3.6.1 Scene Exploration Evaluation Metrics

Regarding the scene exploration task, the priority is to conceive a unified metric that permits the assessment of the performance in spite of the active perception method that is applied. For this reason, and given that the main objective of this task is to accurately map the maximum amount of semantic content available in the scene [68] we consider a recall-like metric (2.5) that measures what fraction of the visual scene has already been correctly mapped in (3.8) at a given timestamp. As stated in Section 2.3, recall measures the ability of a model to detect all ground-truth information. Therefore, the performance of the methods, presented in Section 3.4.1, is assessed with the success rate metric, expressed for time t as

$$\text{Success Rate} = \frac{1}{X \times Y} \sum_{y_m=1}^{Y} \sum_{x_m=1}^{X} \Omega\left[C^{x_m,y_m} \mid \mathbf{p}_t^{x_m,y_m}\right] , \tag{3.26}$$

assuming that all (x_m, y_m) cells contain a ground-truth object, which consists of the ratio between the number of map cells where the highest categorical score (3.9) fits the class, C^{x_m,y_m}, of an overlapping ground-truth object (TP), and the number of cells containing ground-truth objects ($\mathrm{TP} + \mathrm{FN}$). Hence, the Ω function, which identifies whether C^{x_m,y_m} is maximally represented in (3.1) or not, is defined as

$$\Omega\left[C^{x_m,y_m} \mid \mathbf{p}_t^{x_m,y_m}\right] = \begin{cases} 1, & C^{x_m,y_m} = \underset{k}{\operatorname{argmax}} \left\{ p_{t,k}^{x_m,y_m} \right\} \\ 0, & \text{otherwise} \end{cases} . \tag{3.27}$$

Moreover, considering an experimental dataset with multiple images, averaging (3.26) over all the instances of that dataset, for each gaze fixation, would be of added value for the evaluation of the task.

3.6.2 Visual Search Evaluation Metrics

When it comes to the visual search task [17], the best metric of evaluation must intuitively revolve between the speed of the search method, i.e. how many iterations does the model need, on average, to find a given target object. Similarly to the approach applied in the evaluation of the IVSN [29], a sort of an *Oracle* decides whether the task has been completed or not, essentially by confirming its success and terminating the search process every time the gaze is directed towards the region where the target is located. Given the spatial division of the image into a grid of cells, adopted in this methodology, it is considered that the goal has been reached after directing the focal point to a target-containing cell. The semantic content comprised in a cell is determined on the basis of the available ground-truth information. When it comes to the task of searching for specific target objects in a complex scene containing multiple entities of different natures, cumulative performance [29] is considered a significant metric when comparing the results of visual search experiments. Therefore, the success is evaluated through the ratio between the number of images where the target object has already been located over the total number of images in the experimental dataset [16], i.e. the cumulative performance achieved at each fixation.

4

Experimental Setup

Contents

The objective of Chapter 4 is to detail the training process for the foveal observation model and to list generic task-oriented constraints and conditions, arguing that some of them are critical for the global experimental setup, according to the illustration presented in Figure 3.1 in the form of a block diagram, given the nature of the tasks at hand. The validity of the foveal observation model has been tested in a previous study [3], and, for this reason, the content presented in this section is limited to a simple description of the training protocol partnered with some data-related statistics.

4.1 Foveal Observation Model Training Procedure

When implementing an artificial humanoid visual system, it is always necessary to consider the nature of the visual field [5] and the irregularities associated with it. As previously suggested in Chapter 3, more specifically in Section 3.2, the idea behind the foveal observation mechanism is to modulate the uncertainty imposed by the artificial foveal system on the object detection model, namely YOLOv3 [1].

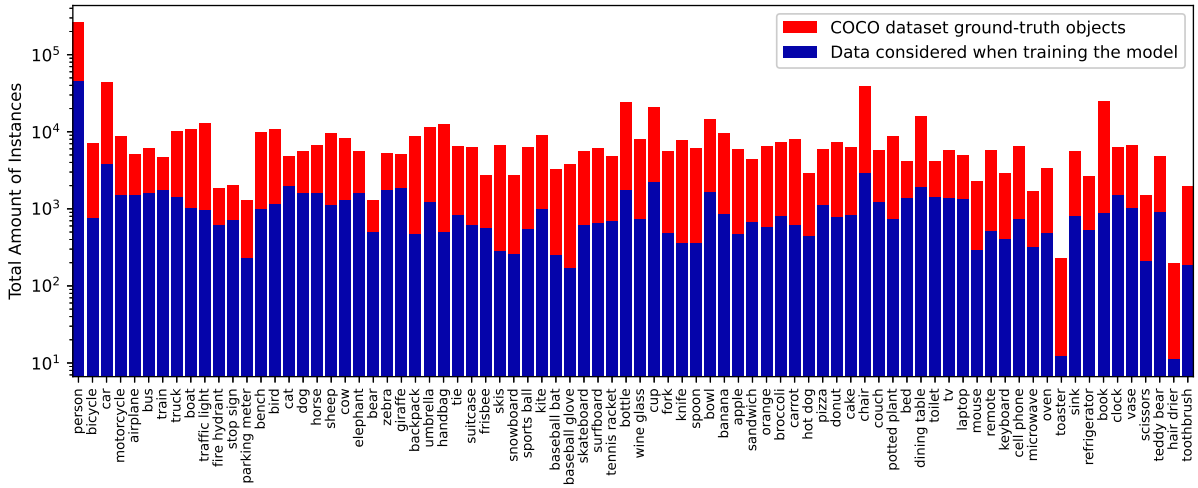


Figure 4.1: Comparison between the amount of ground-truth objects available in the COCO 2017 training-set [16] and the number of filtered YOLOv3 detections during the training phase of the foveal observation model, represented by two histograms and divided according to the respective classes.

Attending to the spatial distribution of the visual field, which consists of a radial increase of the blur level that accompanies a distancing from the center of the foveal region towards the peripheries, it is easy to conclude that there is a correlation between the relative position of an object to the focal point and the uncertainty level associated with its detection. Therefore, the foveal correction mechanism is designed to take advantage of the relative distance between the objects and the focal point to account for the uncertainty inherent to the classifier outputs (3.3), which are presented in Section 3.2.1.

The outstanding advantage of the foveal observation model is that having an object detection model pre-trained with Cartesian images, it is much faster to fit the foveal observation mechanism that models

the uncertainty related to the regional irregularities of the visual field than to retrain the object detection model with a foveal image dataset. Taking advantage of the publicly available YOLOv3 checkpoints, pre-trained on COCO dataset, the foveal observation model training process follows the steps presented:

- All 118287 images from the COCO 2017 training-set are artificially foveated at a randomly selected focal point. Each image serves as input to the artificial foveal system, with the standard deviation parameter on the exponential kernel set as $f_0 = 100$, which defines the size of the foveal region.
- The YOLOv3 object detection algorithm receives every single foveated image as input, filtering the output detections with a confidence score threshold of 0.25 and using a default threshold of 0.45 for the Non-Maximum Suppression (NMS) algorithm.
- Making use of the COCO 2017 training annotations, which contain the ground-truth information of all images, every detection whose bounding-box Intersection over Union (IoU) score with a ground-truth object bounding-box is above the 0.30 threshold is assigned to that object's class.
- Distance is discretized in the form of 7 uniform levels, extending from the center of the fovea to a maximum distance of 514 pixels. Following the approach described in Section 2.4.2, a Dirichlet distribution is then estimated for each class and distance level to the center of the foveal region.

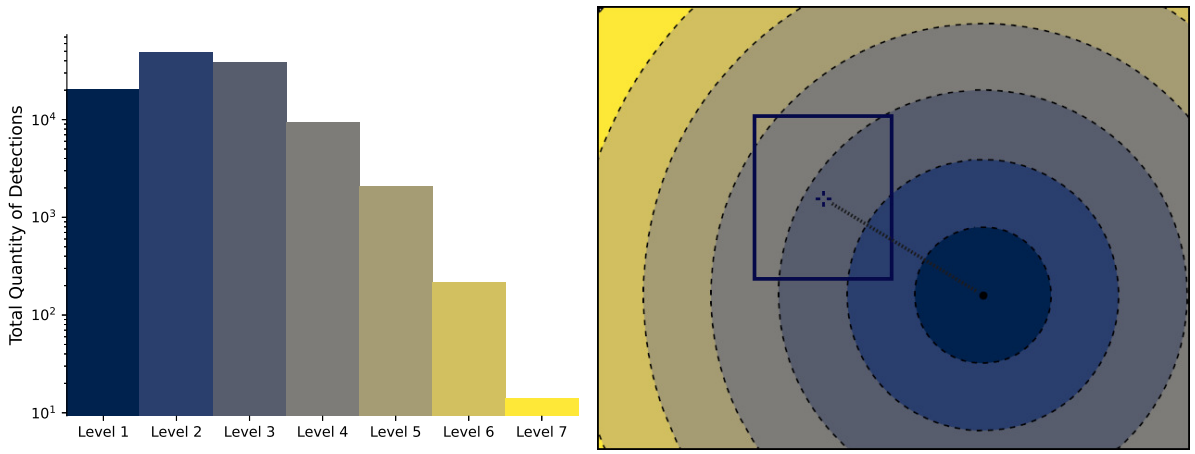
Regarding the foveal observation model training procedure, when considering typical bounding-box qualitative evaluation metrics, namely the mean Average Precision (mAP), it is acceptable to use a standard IoU threshold between 50% and 95% when considering a match between two overlapping boxes. Taking into account the distortion imposed on the visual field by the artificial foveal system, specifically on the peripheral regions, the output bounding-boxes generated by YOLOv3 related to objects located in those regions are consequently less accurate, resulting in the need to lower the IoU threshold while training the foveal correction model, following the methodology presented in Section 3.2.2.

A detection is associated with all the classes of ground-truth objects with which it overlaps. Figure 4.1 presents a comparison between the number of ground-truth objects from the COCO 2017 train dataset and the number of detections associated with each class. The correlation coefficient between the Probability Density Functions (PDFs) [66] associated with two histograms , P and Q, expressed as

$$r\left(P,Q\right) = \frac{\sum_{k=1}^{K}(P_k - \bar{P})(Q_k - \bar{Q})}{\sqrt{\sum_{k=1}^{K}(P_k - \bar{P})^2 \sum_{k=1}^{K}(Q_k - \bar{Q})^2}}, \tag{4.1}$$

is one common metric used to measure similarity, ranging from 0 to 1, where K corresponds to the number of bins (i.e. the number of classes) and $\bar{P}, \bar{Q}$ the respective mean values. To evaluate the similarity between the distribution of the classes, the normalized data of the histograms presented in

Figure 4.1 has been applied to the expression (4.1) resulting in a correlation score of 98.45%, after converting the coefficient into a percentage. The proximity to the absolute correlation reference value leads to the conclusion that YOLO has been able to effectively capture a balanced and evenly distributed representation of the classes contemplated in the COCO dataset, according to their relative abundance.



(a) Detections per Focal Distance Level

(b) Illustration of the 7 Distance Levels

Figure 4.2: Histogram (a) of the YOLOv3 detections distribution per level during foveal correction model training. Illustration (b) of the spatial distribution of the distance levels in reference to the center of the fovea.

Accounting for the radial blur distribution in the humanoid visual field, each matching detection is associated with one of the 7 uniform circular-shaped distance levels. As explained in Figure 4.2(b), the distance level to which a detection belongs is established after computing the distance between the center of its bounding-box and the focal point. The distance is computed by applying the relative coordinates of the center of the detection's bounding-box in relation to the center of the fovea to the expression (3.5), where the parameters σ_x and σ_y of the weight matrix (Σ) are equal to f_0, which is the parameter that defines the size of the foveal region. During the training process, only a few detections (about a dozen) were discarded because their focal distance exceeded the maximum threshold value.

Finally, after applying the 119734 filtered and partitioned detection scores to the Dirichlet Fast Fit algorithm [51], an 80x7 matrix (80 classes per 7 foveal distance levels), containing different Dirichlet distributions, defines what is called the foveal observation model. As explained in Section 3.2.2, each Dirichlet distribution consists of an array $\boldsymbol{\alpha}_{k,d}$ (3.4) that stores 80 alpha parameters, one per class.

When analyzing Figures 4.1 and 4.2(a), one may notice that for some classes (e.g. 'toaster' and 'hair dryer') it may not be possible to compute the alpha parameters for the farthest focal distance levels (e.g. levels 6 and 7) due to lack of data. In such situations, a flat Dirichlet distribution [46], consisting of an array of alpha parameters that are all equal to the unit value $(\alpha_{k,d,i} = 1, \forall\, i = 1, \ldots, K)$, is assigned to that class k and that distance level d on the matrix of distributions, interpreted as a non-informative prior.

4.2 Deep Object Detection Model

As already established, in order to generate object predictions, it is selected the YOLOv3 [1] model, which is a deep object detector that follows the one-step framework, based on global regression/classification. The reason behind the selection of YOLOv3 as the deep object detection model is that, although not the latest version, YOLOv3 remains a fast and accurate model for object detection that uses Darknet-53 [1] as its backbone CNN architecture. Despite not providing overwhelming results on the COCO mAP between the 0.5 and 0.95 IoU thresholds, YOLOv3 delivers a very convincing performance on the original mAP metric (2.8), with a 0.5 IoU threshold. The pre-trained checkpoints presented in Table 4.1 are compatible with a more recent version of YOLO, namely YOLOv5, hence the attributed nomenclature.

Table 4.1: YOLOv3 checkpoints pre-trained on COCO 2017 training-set. All checkpoints are trained to 300 epochs with default settings. The mAP values are for a single-scale model on COCO 2017 validation dataset.

Models*	mAP in COCO val2017		CPU Speed	Total Parameters
	50-95	50	milliseconds (ms)	millions (M)
YOLOv5n	28,0	45,7	45	1.9
YOLOv5s	37,4	56,8	98	7.2
YOLOv5m	45,4	64,1	224	21.2
YOLOv5l	49,0	67,3	430	46.5
YOLOv5x	50,7	68,9	766	86.7

*github.com/ultralytics/yolov3.git

An object prediction outputted by YOLOv3 consists of a bounding-box (defined by the coordinates of the top-left and bottom-right corners), a normalized classifier vector containing categorical probability scores (3.3), and an objectness score $P(\text{Obj}|\mathcal{B}_{t,l})$. Although not considered by all object detection models [18], the objectness score helps to determine whether a bounding-box, for instance, predicted by YOLOv3 contains an object of interest by representing the confidence in the presence of objects within its spatial limits. Low objectness scores can be used to filter out imprecise detections, which results in a reduction of false positives and an improvement of the accuracy of the system. The confidence score

$$\text{Conf}_{t,l} = \max_{k}\left\{s_{t,l,k}\right\} \times P\left(\text{Obj}|\mathcal{B}_{t,l}\right) , \quad k = \{1, \ldots, K\} , \tag{4.2}$$

associated with an object detection outputted by YOLOv3, is used to filter some detections, based on some pre-defined threshold value suitable for the context in which this particular deep detection model is being applied. Considering a real-time implementation of the methodology illustrated in Figure 3.1 in a humanoid robotic system, it is important to account for both the time and memory resources consumed by YOLOv3, given the global computational cost involved in completing the visual tasks. In order to prioritize the speed of execution, the YOLOv5s checkpoints, pre-trained in the COCO 2017 dataset and enumerated in Table 4.1, are utilized by the object detection model during the experimental phase.

4.3 Experimental Apparatus

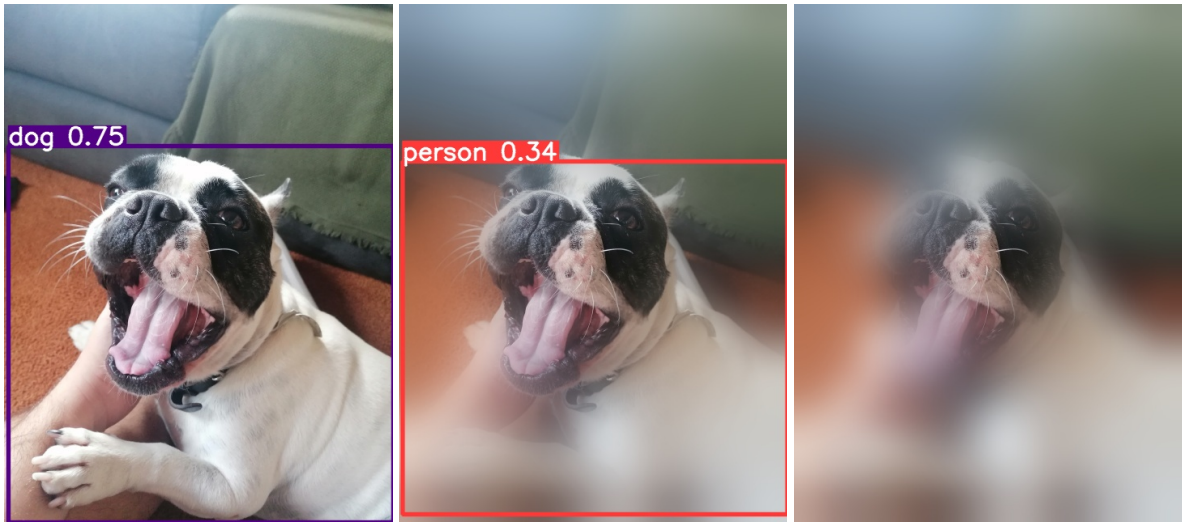
The conglomerate in which the models that comprise the setup for the experiments are arranged, presented in the form of an interrelated block diagram in Figure 3.1, must comply with certain regulations that govern the experimental protocol. These regulations require the definition of the parameters of the constituent models and the experimental data set, established in accordance with the following criteria:

- The experimental test set consists of 300 images selected from the COCO 2017 validation dataset. The first criterion of acceptability is that both dimensions of the images (height and width) must either equal or surpass the amount of 480 pixels each.
- Each chosen image must be comprised of at least 8 distinct objects, instances of the 80 classes represented in COCO dataset that are not deprecated by YOLO, that compose the set of ground-truth objects of that particular image.
- Concerning the application of the artificial foveal system, comprised by a Gaussian pyramid with a total of $K = 5$ levels, the parameter dictating the foveal region's size has been established as $f_0 = 100$, in line with the foveal observation model training procedure.
- The YOLOv3 model, pre-trained on the COCO 2017 training dataset, is used to detect objects in foveal images, filtering the output with a confidence threshold of 0.30 due to the blur imposed on the images and using the default value of 0.45 as the Non-Maximum Suppression (NMS) threshold.
- Each image is mapped into a 10x10 grid of cells. All cells have equal proportions and an object is considered to be contained inside a cell if its bounding-box intersects at least 10% of the cell area.
- One random ground truth object is selected to serve as the target. This means that the class of this particular object is defined as the target class associated with the image. The selected target class can not be contained in more than 20% of the cells that comprise the grid map of the image.
- The initial focal point associated with each image is defined as the center of a pseudo-randomly selected cell. This cell cannot be a target cell, i.e. a cell that contains an instance of the target class associated with the image. It is also ensured that YOLOv3 is able to make at least one object prediction when the foveal region is centered in the exact same cell.
- Implementation of the Inhibition of Return (IOR) mechanism, which prevents access to cells that have previously been visited. Regional saliency inhibition is also applied to saliency maps.

From the 5000 images available in the COCO 2017 validation dataset, a small but significant subset of 300 images is selected to form the experimental dataset, according to the characteristics and parameters that make the images suitable for both visual search and scene exploration contexts. This involves

the selection of images that contain more than just a few detectable objects, otherwise, the amount of semantic information available to be gathered by the object detection model would be very limited. The introduction of bias in the experimental setup is also avoided since the foveal observation model is trained exclusively with content from the COCO 2017 training-set. Consequently, it is guaranteed that the foveal correction mechanism had no prior access to the data available on the experimental dataset.

It is also crucial to consider the limitations imposed by the artificial foveal model on the object detection model. Recall that YOLOv3 model is trained directly on the Cartesian images that comprise the COCO training-set [3], without being submitted to any foveation process. Therefore, since this methodology heavily relies on the ability to detect as many objects as possible, regardless of their size and shape, defining a fovea that is large enough to aggregate the maximum amount of objects, while not deviating much from the structure of the human visual field, appears to be seemingly advantageous.



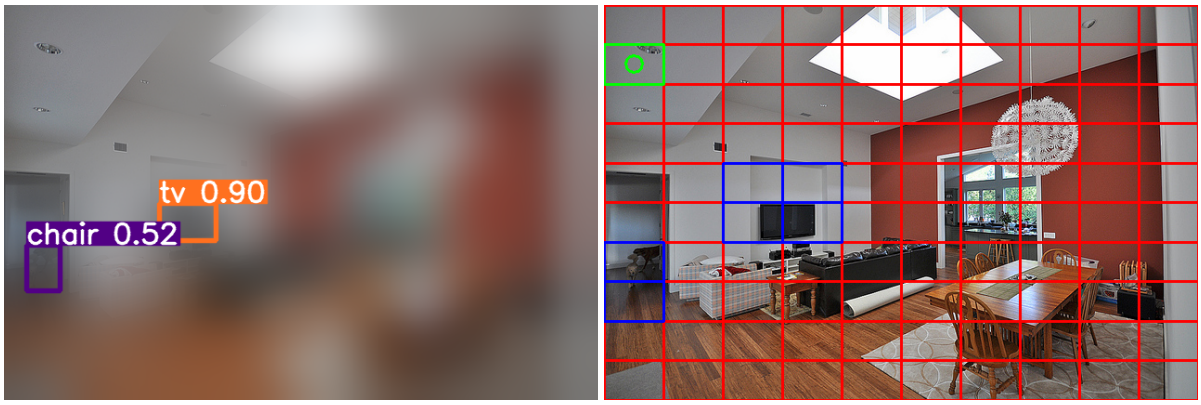
(a) YOLOv3 on Cartesian **(b)** YOLOv3 on Foveal ($f_0 = 100$) **(c)** YOLOv3 on Foveal ($f_0 = 50$)

Figure 4.3: Impact of the limitations imposed by the parameters of the artificial foveal system (more specifically the dimensions of the foveal region) on the accuracy of object detection with YOLOv3.

The importance of this condition is connected to the fact that in the fovea, i.e. the region that surrounds the focal point, the degree of visibility is very similar to the observed in regular Cartesian images. Consequently, this not only facilitates the object proposal mechanism in YOLO, as supported in Figure 4.3, but also possibly leads to an increase in the number of detections, as it is very difficult to detect objects of large proportions in a tiny focal region. To comply with the implications of this limitation, the standard deviation of the exponential kernel (f_0), which is part of the artificial foveal system, must be properly tuned for both the training of the foveal observation model, described in Section 4.1, and during the experimental phase, with equal parameter values. To ensure the presence and influence of the peripheral region on the visual field, the dimensions of the selected images must necessarily exceed the

defined threshold of 480 pixels, to faithfully simulate the human visual field and better approximate the search and exploration tasks in a real scenario. This grants that the foveal region does not occupy the vast majority of the pixels and that there is a vast blurred area (peripheral region) where to direct the gaze when exploring or searching for objects, retaining previously unknown semantic information related to the objects that can be present in those regions.

Due to the blur and distortion imposed on images by the foveal system, not only the accuracy of the bounding-boxes but also the object confidence scores of the detections decrease when compared with detections on Cartesian images. To account for peripheral distortion, the confidence threshold is set to 0.30, similar to the value set when training the foveal observation model. It could be interesting to consider reducing the value of this threshold to maximize the number of considered detections (e.g. the situation described in Figure 4.3), but this would always be connected to a trade-off between quantity and quality, which is not considered a central issue given the broad context of this work.



(a) YOLOv3 detections on Foveal Image

(b) Semantic Map's 10x10 Grid of Cells

Figure 4.4: Example of a simulated visual field (a) and the respective (red-colored) grid of cells (b) that divides the scene's semantic map, with the focal point set on a cell located in the top-left corner. The green-colored cell that is visible on the 10x10 grid of cells (b) marks the center of the fovea and the cells that are updated because of the two affiliated detections, generated by YOLOv3 (a), are highlighted in blue.

Regarding the elimination of overlapping bounding-boxes with NMS [40], the typical threshold values, applied in the detection of generic objects, range between 0.40 and 0.50, therefore 0.45 is the default NMS threshold value for YOLOv3. The definition of such parameters usually depends on the context of the application and the particular characteristics of the data involved. By dividing the image into a cell grid, as illustrated in Figure 4.4, it is possible to map the semantic information associated with multiple detections to their respective regions, defined as rectangular cells smaller than the foveal area. As a result, a less localized (with respect to the pixel level) but still spatially organized distribution of the collected information is made available through successive updates of the map. A uniform Dirichlet distribution ($\beta_k^0 = 0.5$ for $k = 0, \ldots, K$) is assigned to each cell, following the conclusions taken

by Kaplan [62] and applied by Dias [3], ensuring initial equiprobability between all classes, including the background class, throughout the map. One way to ensure that, initially, all experiments collect information from the same region of the visual field is to define one starting cell, associated with each image, on which the initial focal point is centered. The advantage of this condition is that it ensures a more fair comparison between different active perception methods, especially between those that depend on semantic information to build the cell map. Based on the nature of the visual search task, it seems much more interesting to define these initial cells in a pseudo-random fashion rather than by a purely random approach. Therefore, the selected starting cell must not contain an instance of the target class.

Concerning the choice of the initial cell, it is also granted that YOLOv3 can make at least one object prediction when the focal point is set in the same cell. In some cases, it is likely that either no relevant object is visible in the fovea or that YOLOv3 is simply unable to detect any object. In such a situation, any active perception model that depends on the semantic information gathered on the (3.8) map will have no data available to make an informed decision regarding the next best focal point. In the occurrence of such an event, the best solution would probably be a random saccade, which could also lead to a no-detection zone. Given the limitation on the number of iterations (i.e. gaze shifts) for each image of the experimental set and the desire to purely evaluate the influence of semantic information usage when completing the tasks, the imposition of this constraint seems certainly unavoidable.

When considering the usage of information extracted from a visual field saliency map to select the next best focal point, it is also important to consider the incorporation of an IOR mechanism [17] to avoid stuck states. For instance, when considering the visual search task, this condition proves to be quite critical because, although there might be one single most salient region on the entire image, if this region does not contain the target object, one desires to search on the second most salient region [29], and so on. Without inhibition of previously visited regions, the model tends to get stuck in the most salient region, leading to an unintended outcome, which is contrary to the objective of the experiment. In addition to the map that builds up information step-by-step, through sequential updates, during the experiments, another map is created to capture the ground-truth information available on the scene, allowing for a more accurate evaluation of the task's performance by different active perception methodological approaches. This other map stores the classes of objects contained in each cell.

Finally, regarding the adaptation of the saliency-based, VOCUS2 [15], it is considered a pixel intensity threshold of 0.75 (in relation to the most salient cell), with which the most conspicuous pixels are identified, in order to identify RoIs through regional association to the cells of the grid. Among the experimental protocol, there are other task-specific constraints, like the number of iterations simulated, i.e. the maximum number of gaze shifts towards another focal point, which will be discussed in due course.

5

Experiments & Results

Contents

A succinct description of the experiments performed regarding the two tasks addressed explicitly in this body of work, namely scene exploration [26] and visual search [17], is provided in Chapter 5, together with the most relevant results that may allow concluding about the possible advantages or disadvantages of using semantic information when scanning the visual field with a pre-defined objective.

5.1 Experimental Overview

In order to plan the experimental endeavors, it was conceived a set of research questions that generated a list of experimental objectives. These experimental objectives have been established in accordance with the goals of both scene exploration and visual search tasks, which can be directly linked with the intricacies of the methodology presented in Chapter 3. The objectives are described in the following list:

- Compare the accuracies of the semantic-based active perception methods [3] and the saliency-based model [15], when it comes to completing both scene exploration and visual search tasks.
- Understand whether the foveal observation model positively impacts the performance of the active perception model when completing the visual tasks, in comparison to the non-calibrated approach.
- Given that this is the first attempt at using the semantic-based active perception approach to perform visual target search, it is important to establish whether the methodology [3] is suitable for the task, through a comparison between the predictive (3.25) and non-predictive (3.24) approaches.

Having in sight the enunciated experimental objectives, with regard to the experiments related to the scene exploration task [26], the following active perception methods are tested and posteriorly evaluated:

- Incorporation of the KL divergence metric (3.18) with the acquisition function (3.21).
- Incorporation of the classification negentropy metric (3.20) with the acquisition function (3.21).
- Incorporation of the two-peak difference metric (3.22) with the acquisition function (3.23).
- A most salient cell selection, from the saliency map generated by VOCUS2 [15] at each fixation.
- A random gaze selection algorithm, that picks the next focal cell on an aleatoric basis.

The semantic-based active perception approaches, i.e the ones that consider (3.18), (3.20), and (3.22), follow the full methodological approach defined in Section 3.4. Moreover, each semantic-based approach is tested in two distinct scenarios. These active perception approaches are tested with a semantic map that is sequentially updated either with foveal calibration (3.12) or without any calibration at all (3.10), directly with the classification scores outputted by YOLOv3, as represented by the straight and dotted line paths that flow from the object detection block in Figure 3.1, respectively. As explained

in Section 4.3, after each iteration, i.e. each time the gaze is shifted, the IOR mechanism blocks the cell that is being focused on the moment, preventing it from being accessed in further iterations, corroborating the schema presented in Figure 3.1. The IOR mechanism is applied in all the experiments, regardless of the active perception method being applied, both for scene exploration and visual search. Furthermore, regarding the visual search task [17], the tested and assessed methods are as follows:

- Non-predictive active perception approach (3.24), considering the knowledge accumulated in (3.8).
- Predictive active perception approach (3.25), considering the expected knowledge from (3.14).
- A most salient cell selection, that also assesses the saliency map generated by VOCUS2 [15].
- A random gaze selection algorithm, that also selects the next focal cell on an aleatoric basis.

Regarding the statistical representativity of experiments that follow the protocol established in Chapter 4, it is important to consider the influence of the starting cells on the exploration results and the variability that can derive from the initial focal point. To account for the influence of the initial focal conditions, each experiment is repeated 10 times, and only the mean value of the evaluation metrics is presented together with the associated Standard Error of the Mean (SEM), in the form of bars or bands.

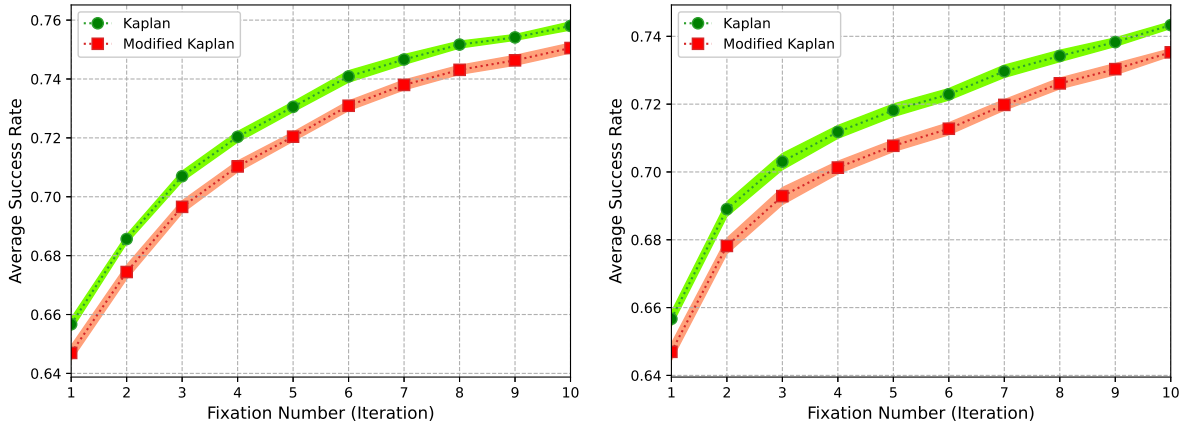
5.2 Scene Exploration

As previously explained in Section 3.6.1, when exploring a visual scenario, the goal of a human is to discover and localize the maximum number of objects displayed on it [69], while performing the minimum amount of actions to obtain a better perception of the visible environment. In the context of visual perception, changing the focal point by performing a saccade is considered an active step [68]. Consequently, the experiments aim to understand to what extent the different perception methods lead to accurate scene mapping when performing a few initial gaze shifts, simulating human scene recognition. For this reason, the number of iterations has been limited to 10 for all the methods involved in the scene exploration experiments, with the average success rate (3.26) being assessed in each single fixation.

5.2.1 Impact of the Foveal Observation Model in the Success Rate

When considering the scene exploration experimental procedure, it is important to first distinguish between the data fusion technique applied to the semantic map used for active perception, where updates are simulated to determine the next best viewpoint, and the technique applied to the semantic map used for the task's success evaluation. Therefore, the level of accuracy detecting objects within each cell can be deduced either from the Kaplan map, which is directly updated with the YOLOv3 classification outputs using the Kaplan rule (3.10), or from the Modified Kaplan map [3], updated with the adapted rule (3.12)

using the calibrated YOLOv3 scores obtained from the foveal observation model. Based on the duality of semantic information mapping, either directly using the outputted scores (3.3) or using the calibrated foveal scores (3.6), it is inevitable that the map used for the next focal point selection influences the sequential evolution of the other. This cross-interference comes from the fact that the detections extracted from a focal point and used to update both maps are a direct consequence of the selected method. Therefore, the usage of one map may lead to a more accurate description of the scene on the other.



(a) Random Focal Point

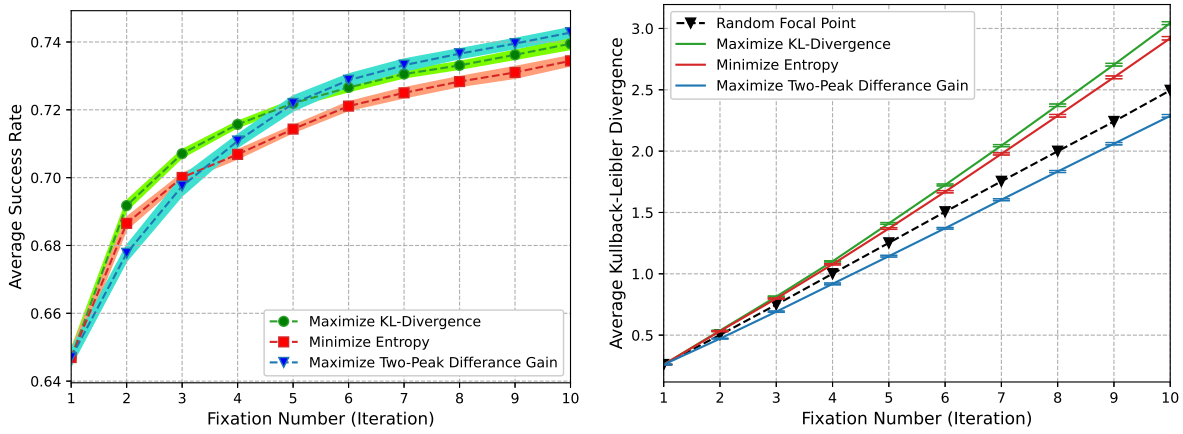
(b) Most Salient Region

Figure 5.1: Average success rates observed during the scene exploration experiments, when defining the next best viewpoint (a) randomly and (b) as the most salient cell from the whole map generated by VOCUS2 [15].

Figure 5.1 highlights the success rate results obtained when inferred from the map directly updated with the YOLOv3 scores (3.10), i.e. the Kaplan map, and the map that is updated with the corrected foveal scores (3.12), i.e. the Modified Kaplan. In addition to the two gaze shift decision approaches presented in Figure 5.1, there are also other figures in Appendix B obtained from experimenting with other active perception methods that corroborate the observable results. An analysis of the experimental results naturally leads to the conclusion that mapping the scene directly with the YOLOv3 scores (Kaplan) results in a more accurate categorical representation of the available semantic information compared to mapping with the corrected foveal scores (Modified Kaplan), regardless of the method being applied in the experiment. The fact that these results shed light on concerns about which map better explains the semantic substrate of an object-filled visual field does not undermine the utility of the foveal observation model, as it is still foundational for the active perception methodology. The predictive step of updating the map with the expected value of the categorical distribution of each cell, expressed in (3.17), directly depends on the local and categorical probability distributions included in the foveal observation model.

5.2.2 Active Gaze Control with Different Uncertainty Measures

After establishing a comparison between the two maps that make up the experimental methodology concerning task evaluation, the attention is now turned toward predicting the next best focal point, specifically the comparison between the different acquisition functions [57] regarding the ability to generate an exploration path that maximizes the amount of semantic knowledge gathered from the visual field.



(a) Success Rate

(b) Kullback-Leibler (KL) Divergence

(c) Classification Entropy

(d) Two-Peak Difference

Figure 5.2: Comparison of the mean of the average values of success rate (a), KL divergence (b), entropy (c) and two-peak difference (d) with different active perception techniques that consider the information accumulated in the Modified Kaplan map, updated through (3.12) with foveal calibrated scores (3.6). The means and respective SEM bands are obtained from 10 repetitions, with different initial focal cells.

As explained in Section 3.4, acquisition functions fundamentally serve as uncertainty quantifiers that facilitate inferring the level of confusion in different regions of the map. Figure 5.2(a) presents the success rates (3.26) achieved when minimizing the expected uncertainty distributed throughout the map

with three different information-theoretic metrics, namely the KL divergence (3.18), the classification entropy which is simply the symmetrical value of the negentropy (3.20) (i.e. the lower the entropy the lower the confusion level presented by the cell) and the absolute probability difference (3.22) between the two classes that present the largest expected class posterior probabilities (3.1). In order to account for the effect of the foveal observation model, presented in Section 3.2.2, the results presented in Figure 5.2 involve information extracted from the semantic map updated with the Modified Kaplan rule (3.12). From the results displayed in Figure 5.2 the maximization of the Kullback-Leibler divergence is elected as the best-performing active perception metric among those considered. The KL divergence metric (3.18) consistently outperforms the entropy metric (3.20) and, in relation to the two-peak difference gain (3.22), presents a better trade-off between the amount of correct information gathered in the first iterations and the long-term performance of the active perception model. Moreover, the results presented in Figures 5.2(b), 5.2(c), and 5.2(d) corroborate that the KL-based method excels all the others, in their own metrics.

5.2.3 Comparison with Traditional Saliency Model

In a second instance, the proposed semantic-based model [3] is compared with both a random gaze selection algorithm, which randomly selects the next gaze direction, and VOCUS2 [15], the elected saliency-based method. Furthermore, the usage of calibrated (3.6) and non-calibrated (3.3) scores in the map update stage is assessed, using the Modified Kaplan (3.12) and Kaplan rules (3.10), respectively.

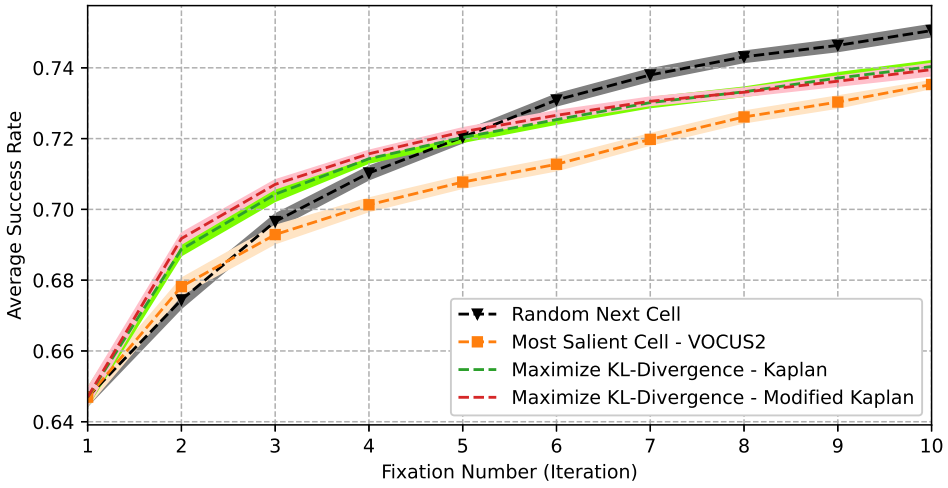


Figure 5.3: Mean values of the average success rate and the respective standard error (SEM) bands, for the semantic-based active perception approach that considers the probability of improving (3.21) the KL divergence metric (3.18) using either the Kaplan or Modified Kaplan updated maps as sources of information, in comparison with both the random gaze selection and VOCUS2 [15] (saliency model) results.

In Figure 5.3 are presented the obtained average success rates (3.26). By maximizing the KL divergence, the methodology proposed by Dias et al. [3] not only outperforms the saliency-based active

exploration [15] but also initially leverages the amount of semantic content that is correctly mapped even in the face of a random selection. Moreover, there is no significant upside in considering the calibrated scores (3.6), fused in the Modified Kaplan map, rather than the raw scores (3.3), fused in the Kaplan map. After about 4 saccades, the random algorithm eventually outperforms all the other methods.

5.2.4 Computational Cost of the Active Exploration Algorithms

Finally, from the data presented in Table 5.1, it is possible to draw the conclusion that the implementation of the semantic-based model [3], described in Chapter 3, reveals itself to be very costly. The disparity between the time-cost values obtained with VOCUS2 [15] and the proposed methodology [3] comes from overhead imposed by a quadratic complexity on the update process upon the exploration algorithm.

Table 5.1: Comparison between the relative computational costs of the implemented version of the methodology proposed by Dias et al. [3], as well as the adaption of the VOCUS2 [15] model and the random gaze selection algorithm. The results are presented either in seconds (s) or milliseconds (ms), per iteration.

	Semantic [3]			Saliency [15]	Random
	KL-Divergence	Entropy	Two-Peak Difference	VOCUS2	
Kaplan	9.851 s	9.770 s	**8.971 s**	**0,462 s**	0,017 ms
Modified Kaplan	9,907 s	9,833 s	**8,981 s**		

The fact that for every single cell, out of the $X \times Y$ grid, a total of $X \times Y$ updates must be computed, to simulate a gaze shift toward each (x'_m, y'_m) cell, appears to be quite inefficient when compared to the cost of generating a saliency map. Nevertheless, the proposed methodology is still able to enhance the amount of accurately mapped information in the first iterations [3], which is a major upside for the model.

5.3 Visual Search

In the work developed by Dias et al. [3] the benefits of using semantic information, retrieved from an object detection model to map objects displayed in a scene, are considered and tested. This means that the active perception component of the experimental apparatus is designed to sequentially select new focal points in a way that aims at extracting as much indiscriminate semantic information (i.e. non-class specific information) as possible to constantly reduce the global level of confusion on the whole map.

The objective of the experiments presented in this section is to figure out whether, by adapting the active perception metrics to target a specific class, it is possible to navigate a visual scene to seek an instance of a pre-defined class [17], using the approach described in Figure 3.1. Essentially, the idea is to understand how adaptable the design of the model to a different goal within the human visual framework is, discerning whether the usage of semantic information is advantageous when aiming at efficiently finding a categorically specific object rather than simply describing the semantic content of

the scene. To obtain precise results, each experiment was repeated 10 times with different initial cells to accommodate any potential impact of the starting focal point. Since the goal is to understand how many gaze shifts are necessary to achieve success, all experiments were terminated after 30 iterations, ensuring an extensive operation margin for all search algorithms. In the context of visual search, success is achieved when the model shifts the gaze toward a cell that contains an instance of the targeted class.

5.3.1 Comparison between Predictive and Non-Predictive approaches

Consider now the first goal of the visual search experiments, which consists of comparing the effectiveness of predictive active perception (3.25), with simulated updates, against relying directly on raw information gathered from multiple object detections and stored on the semantic maps to identify the next best focal point (3.24), without simulation of updates. Moreover, the aim is also directed to the investigation of potential benefits that proceed from the implementation of the foveal calibration, determining whether it can facilitate efficient identification of cells that contain instances of the target class.

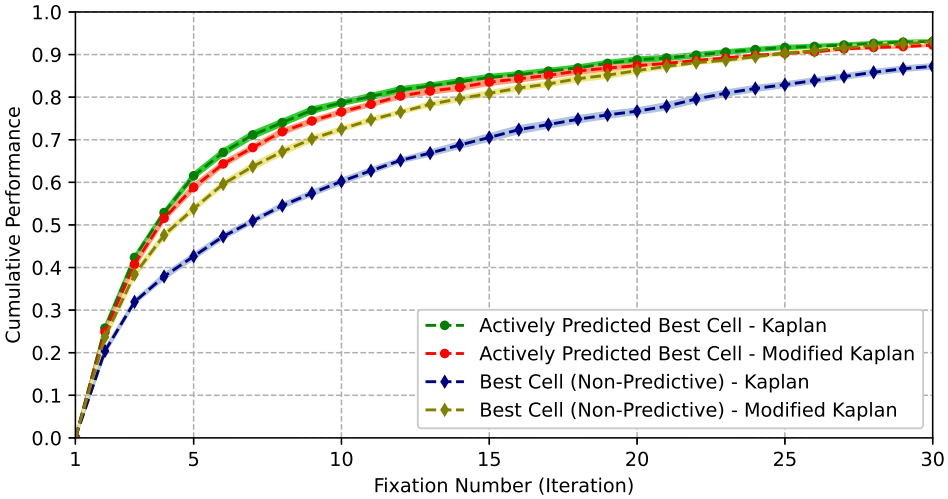


Figure 5.4: Comparison between the mean values of the cumulative performance and the respective standard error bands, obtained using different search metrics. These results were observed after completing 10 experiments, each starting in a different focal point, using predictive and non-predictive approaches that are realized using the semantic information available either in the Kaplan or Modified Kaplan maps.

Given that, when performing the visual search task, the objective is to localize an object rather than map the general semantic content of the scene, according to the perceptive approach described in Section 3.4.2, the confidence measurement functions (3.24) and (3.25) only take into consideration the target's categorical probability in each cell, individually, ignoring the information available in adjacent cells. Nevertheless, despite the greedy approach of the methods with which the experiments were conducted, the algorithm's complexity is reduced from quadratic to linear in comparison to the scene exploration perceptive approach, because each simulation consists only of updating a single cell, in the

predictive approach, rather than the whole map for each possible next cell that can be fixed.

Two main conclusions arise from the results portrayed in Figure 5.4. The first conclusion emerges from the non-predictive search algorithms that select the next fixation point as the center of the most promising cell just by observing the categorical probability of the target class at the current state of a map, after being updated with the most recent detections, extracted during the latest iteration. After applying this algorithm 10 times, starting in different cells, either using the Kaplan map, updated directly with the raw YOLOv3 score vectors, or using the Modified Kaplan map, updated with the calibrated foveal scores, it is quite clear that the latter leads to a higher performance since it can successfully find more objects in the same number of iterations when considering any of the first 30 fixations.

Table 5.2: Mean values of the cumulative performance observed at different experimental stages along with the respective maximum SEMs and computational costs per iteration, expressed in milliseconds. The values are extracted following the results presented in Figure 5.4, contemplating the semantic-based search with predictive and non-predictive approaches using both Kaplan and Modified Kaplan maps.

	Predictive		Non-Predictive	
	Kaplan	Modified Kaplan	Kaplan	Modified Kaplan
Cumulative Performance @ 5	**61,53**%	58,83%	42,63%	**58,83%**
Cumulative Performance @ 15	**84,60**%	83,53%	70,53%	**83,53%**
Cumulative Performance @ 30	**93,10**%	92,23%	87,23%	**92,23%**
Maximum SEM	0,772%	0,820%	0,845%	0,718%
Computational Cost	77,81 ms	83,76 ms	0,144 ms	0,133 ms

As discussed in the work of Dias et al. [3], the foveal correction model does not produce less accurate score vectors when compared to those produced by YOLOv3, demonstrating the validity of the foveal observation model. However, the level of uncertainty associated with the detections is significantly reduced, especially when considering the detections to which YOLO attributes very low confidence scores, due to the blur and distortion that affect the peripheral regions, where most of these detections are extracted. The classification entropy (3.20) of calibrated foveal scores is generally either very low or just residual [3, 63], filtering, to a certain degree, the amount of uncertainty that could be transmitted to semantic maps through successive updates, through the application of the Kaplan rule.

One possible explanation for these results is that the foveal observation model tends to, on the one hand, heavily reduce the categorical probabilities of the classes that are very unlikely to appear at a certain focal distance level, for a certain YOLOv3 score vector associated with some object detection. On the other hand, the foveal observation model tends to enhance the probabilities of the classes that are typically correlated with the same score vector and distance level, reducing the confusion that is transferred to the semantic map. Through the analysis of the cumulative performance mean values that result from the non-predictive approaches, shown in Table 5.2, the respective SEMs, and attending to the fact that the active perception model relies directly on the information available on the semantic

map, without applying any predictive step, the conclusion that can be consequently inferred is that the application of the foveal correction mechanism, when mapping the semantic information retrieved from an object detection model, leads to a higher level of accuracy when finding instances of a specific class.

The second conclusion that emerges from this set of experiments, based on the results displayed in Figure 5.4, is related to the question of how worth the usage of a predictive approach (3.25) is when considering which cell should be gazed upon in the next saccade. From the values presented in Table 5.2 it is easy to infer that, at least until the 20^{th} iteration, the predictive approaches outperform the non-predictive approaches, regardless of the map considered by the active perception mechanism. Therefore, simulating a saccade toward every cell, through the expected value of its categorical distribution (3.14), using the parameters of the Dirichlet distributions (3.4) that correspond to the first distance level, i.e. the center of the fovea as described in Section 3.4.2, corroborates the increase in efficiency.

It is also clearly noticed that, between the 5^{th} and 10^{th} iterations, the predictive approach that takes advantage of the information available in the Kaplan map slightly outperforms the predictive approach that considers the calibrated semantic information accumulated in the Modified Kaplan map. This advantage may be connected to the fact that the alpha parameters, used to infer the expected value of each cell's probability distribution, have been estimated [51] to, as explained in Section 4.1, fit the raw YOLOv3 semantic data that is used to update the Kaplan map [3], rather than data previously calibrated using the foveal observation model, violating the fundamental logic behind the design of the model.

5.3.2 Comparison with alternative Search approaches

In the context of scene exploration, Dias et al. [3] emphasize the benefits of active perception over randomly traversing the map. Although, from a biological stand, the most relevant comparison would be between the performance of real humans and the performance of the semantic information model, this work aims at making the comparison with a neurological model [9], inspired by the behavior of the early primate visual system. Nevertheless, previous studies [25, 29] have proven that, generally, humans tend to outperform unguided search algorithms. Hence, a random approach is used as a baseline model.

As previously asserted in Section 3.6.2, an *Oracle* [29] is responsible for terminating the search algorithm once the focal point is set upon a target cell, indicating that the search has been concluded because one instance of the target class has been found, and that the main goal has been reached. Despite the broad acceptance criteria for success in visual search, related to the number of target cells, that can occupy up to 20% of the map's cell grid, it is crucial to understand to what extent is the semantic information model able to leverage the level of accuracy with which the task is completed.

Once again, there are two main distinctions to be made, when interpreting the results presented in Figure 5.5. Firstly, it is important to assert whether a semantically guided search outperforms an unguided search approach, which simply consists of randomly moving through the cells of the map.

Secondly, given that, according to the FIT, the bottom-up information registered in feature maps is part of the guiding mechanism that directs the attention toward conspicuous regions [17], there is interest in understanding how well a model like VOCUS2 [15], which mimics the traditional saliency-based cognitive model [9] that complies with the FIT, can deliver in terms of finding a target object by sequentially navigating through the most conspicuous cells. Considering the experimental conditions and constraints described in Section 4.3, recalling that instances of the target class do not cover more than 20% of the cells in the images, the results presented in Figure 5.5 corroborate the probabilistic analysis derived from the chance of randomly selecting a target cell. On average, about 13% of the cells are target cells.

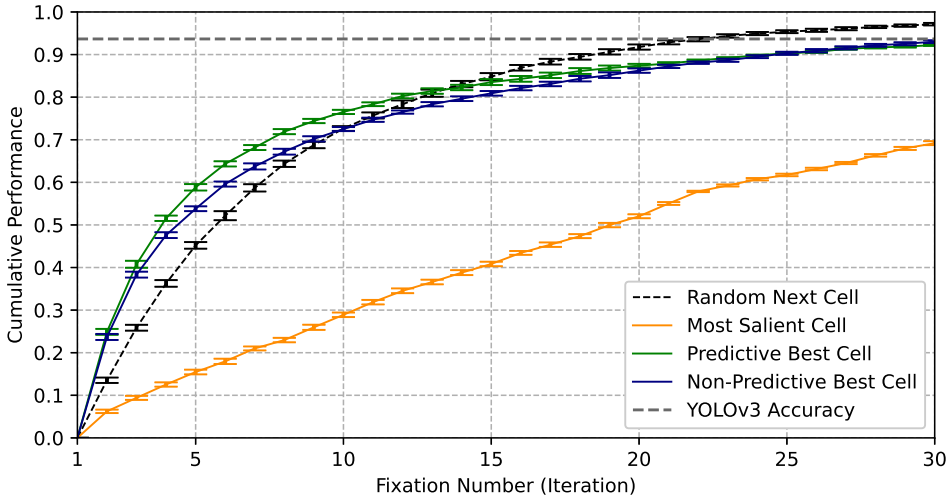


Figure 5.5: Comparison between the mean values of the cumulative performance and the respective SEM bars, obtained using semantic [3] and saliency-based [15] approaches, together with the random search results, observed after completing 10 experiments, starting at different focal points. Both predictive and non-predictive approaches exploit the semantic information fused on the Modified Kaplan map.

From the mean values of the cumulative performance observed in Figure 5.5 it is possible to conclude that the semantic-based predictive search method (3.25) performs better than the unguided random search approach, up to the 13^{th} iteration. It is also interesting to note that the non-predictive approach (3.24), which exploits the YOLOv3 score vectors fused with the Kaplan update rule, takes the upper hand during the first few iterations but eventually falls in the face of the random search algorithm.

Beyond the differences that have been indicated, there is also an offset that is visible in Figure 5.5, related to the cumulative performance values that result from experimenting with the predictive semantic model. To assert whether this offset results from limitations imposed by the object detection model, a simple procedure consisting of feeding all images from the experimental dataset directly to YOLOv3 to understand how efficiently the model can detect at least one single instance of the targeted class. The horizontal dotted line showcased in Figure 5.5 represents the YOLOv3 [1] success ratio in the experimental Cartesian image set, without the application of confidence score filtering to the generated

detections. Virtually, the object detection model considered every generated prediction, regardless of the level of confidence, and if at least one detection possessed a class categorical score that was higher than the score of all the other classes, it was considered that YOLOv3 [1] was able to identify instances of the target class. If a confidence score threshold, set to the same value used during the active perception experiments (30%), was applied during the procedure, the accuracy would drop strictly from 93,67% to 56,33%. Therefore, this limitation might be connected to the reason why random search eventually outperforms the semantic approaches, upon reaching a certain critical number of iterations.

Table 5.3: Comparison between the mean values of the cumulative performance observed at different experimental stages, obtained with the random and saliency-based approaches that can be extracted from the results presented in Figure 5.5, along with the respective maximum SEMs and computational costs per iteration in milliseconds. To allow for a proper comparison, the results obtained with the semantic-based model, using information fused with the Modified Kaplan update rule, displayed in Table 5.2, are also presented.

	Random	Saliency [15]	Semantic [3]	
		VOCUS2	Predictive	Non-Predictive
Cumulative Performance @ 5	45,20%	15,47%	**58,83%**	53,80%
Cumulative Performance @ 15	**84,80%**	40,83%	**84,60%**	80,90%
Cumulative Performance @ 30	**97,13%**	69,20%	93,10%	93,03%
Maximum SEM	1,059%	0,627%	0,820%	0,718%
Computational Cost	0,015 ms	762,3 ms	83,76 ms	**0,133 ms**

Regarding the performance of VOCUS2 in object search, which generates saliency master maps from foveal images and permits the identification of the most conspicuous cells, from the results displayed in Figure 5.5 and extracted to Table 5.3 it is possible to extrapolate that the lack of precision when finding target objects is directly related to the fact that the search algorithm does not consider class-specific information. Therefore, the simulated attention mechanism [15], built on bottom-up saliency information that is extracted from the stimuli at the pixel level, appears to be very inefficient in finding objects in a class-oriented context, in the presence of multiple distractors within the scene's confinements.

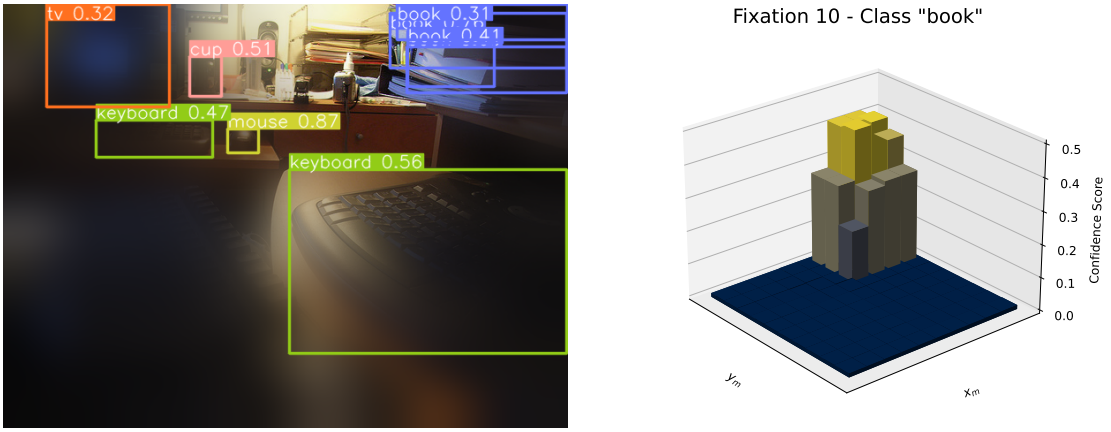
5.3.3 Computational Cost of the Active Search Algorithms

First, from the measured computational times presented in Table 5.2, it is possible to conclude that there is an added cost between the predictive (3.25) and the non-predictive (3.24) search approaches. The leverage of accuracy, proportioned by the predictive approach (3.25), justifies the extra time costs.

Moreover, in a second instance, consider now the time costs associated with the different search approaches, displayed in Table 5.3. The results show that generating a bottom-up saliency map with VOCUS2 [15] is considerably more costly than both predictive and non-predictive semantic-based approaches, proposed in this work. Therefore, the metrics proposed in Section 3.4.2 present themselves to be not only suitable for visual search [17] but also faster than the bottom-up saliency-based [15] model.

5.3.4 Visual Search Example

In his work, Simões [63] presents an example where the semantic-based methodological approach [3] is applied while performing an experiment regarding the scene exploration task. To acquaint the reader with a proper visual exemplification of the target search task [17], a simple example of a class-oriented search is presented, recurring to Appendix C to display the first perceptive iterations, i.e. gaze shifts.



(a) YOLOv3 Object Detections **(b)** Map Confidence Scores

Figure 5.6: Example of a visual search experiment, regarding an image sampled from the COCO 2017 dataset. There are presented the YOLOv3 detections (a) obtained at the 10^{th} focal point as well as the state (b) of the semantic map (3.8), specifically with regard to the confidence scores of the target class ”book”.

In Appendix C are presented the first four iterations that result from the application of the predictive visual search approach (3.25), using information gathered on the semantic map (3.8) which is updated with the foveal calibrated scores (3.6) through the Modified Kaplan rule (3.12). The objective of this particular search experiment is to find the target class ”book”, which is obviously not deprecated by the object detection model YOLOv3 when considering all the relevant classes comprised in the COCO dataset. Notice how in the first iteration, corresponding to an initial random focal point, none of the detections outputted by YOLOv3 are directly correlated with the target class. Nonetheless, after only four iterations, as suggested in Figure C.1, the proposed methodological approach is able to finally direct the gaze toward the region of the field of view where the ground-truth target objects are located. Moreover, notice that, prior to the 4^{th} fixation, the target object was not directly detected by YOLOv3. This supports the idea that by properly detecting distractors and fusing the respective information in (3.8) it is possible to, on the one hand, lower the confidence scores of the target class in the regions where there are no instances of the target class. On the other hand, in the regions where semantic knowledge has not yet been acquired, the confidence is enhanced, representing possible RoIs. This example corroborates the idea that, if meticulously following the pipeline of Figure 3.1, the metrics proposed in Section 3.4.2 are suitable for the visual search task [17] as, in this example, success is achieved very quickly.

6

Conclusion

Contents

In Chapter 6 are highlighted the most valuable results that emerge from the experiments described in Chapter 5. Moreover, some considerations are made about the doors that are opened by the conclusions that flow from the experiments performed in this work. Interesting paths that could be taken next, being either directly or indirectly related to the work developed in this dissertation, are also presented.

6.1 Highlighted Contributions

As initially stated, the aim of this work is to establish how accurately the novel semantic-based model [3] is able to complete the proposed visual tasks, using different active perception metrics, and to perform a qualitative and quantitative comparison with VOCUS2 [15], a bottom-up saliency model. To this end, the active perception methodology proposed by Dias et al. [3] was fully implemented, tested, and adapted in order for it to be suitable for visual search context [17]. To this end, the methodological pipeline illustrated in Figure 3.1 was designed and meticulously followed, in order to fully test the saliency-based [15] and the semantic-based [3] active approaches, with and without foveal calibration [48].

Let us start by considering the results obtained during the experimental phase in regard to the scene exploration experiments. First, it has been shown, through thorough experimentation, that the methodology proposed by Dias et al. [3] is able to accompany the performance of a biologically inspired attention mechanism [9]. Moreover, the ability of the different active perception techniques, presented in Section 5.2, was validated when it comes to accurately mapping the scene's available semantic content. It was also desired to understand whether the Modified Kaplan map, updated with the calibrated scores (3.6), outperforms the Kaplan map, updated directly with the scores (3.3) outputted by YOLOv3 [1].

Regarding visual search, first, it was possible to conclude that by implementing the foveal observation model, which calibrates the scores that are used to update a map (3.8), there is a positive impact on the performance of the semantic model. Furthermore, it is also established that the full predictive approach [3], using the expected value of the probability distribution in each cell to simulate an update, is undoubtedly advantageous, beating the traditional bottom-up saliency model by a substantial margin. Notice that VOCUS2 [15] heavily underperforms in the face of other visual search methods, due to the information considered not being target-oriented. Furthermore, in terms of computational cost, from Table 5.3 it is possible to infer that, in comparison to the cost of generating a full conspicuity map, the active perception mechanism for visual search (3.25) is not only effective but also quite fast. Furthermore, visual search results, presented in Section 5.3.2, support the idea that the human cognitive recognition system does not rely solely on bottom-up features [17], but also on top-down elements and other target similarity principles [36] grounded on the Attentional Engagement Theory (AET). The lack of top-down class-specific features in the model [15, 37] justifies the inaccuracy exhibited when completing the task, in contrast with other recent models (e.g. the IVSN [29]) that incorporate such features.

6.2 Future Work

In this section, the goal is to present a variety of considerations and comments on the methodology, applied in this body of work, as well as the results obtained during the experimental phase. It also discusses the viability of implementing alternative solutions to some underlying problems that emerged while implementing the methodology and indicates possible paths, rooted in the work presented in this document, that can possibly lead to relevant improvements.

As mentioned in Section 3.3.1, it is possible that Dias et al. [3] and, particularly, Simões [63] are missing out on the opportunity to properly model the actual existence of objects of interest in the different regions of the visual field. The proposed approaches do not consider the incorporation of the objectness score when modeling the confidence for the background class, which can be applied in a new formulation

$$s_{t,l,k} = \begin{cases} s_{t,l,k} \times P(\mathrm{Obj}|\mathcal{B}_{t,l}) \ , & k = \{1, \ldots, K\} \\ 1 - P(\mathrm{Obj}|\mathcal{B}_{t,l}) \ , & k = 0 \end{cases} \tag{6.1}$$

that defines the categorical probability of the background class as $P\big(\overline{\mathrm{Obj}}\,|\mathcal{B}_{t,l}\big) = 1 - P(\mathrm{Obj}|\mathcal{B}_{t,l})$, which is the complement of the objectness score. There is no barrier that stops the foveal observation model from also taking advantage of the presented formulation, since replacing $s_{t,l,k}$ with $s'_{t,l,k}$ does not defy any fundamental principles, as it is still granted that $\sum_{k=0}^{K} s'_{t,l,k}$ is a unit-sum. Nonetheless, this simple and intuitive novel approach still demands proof of concept, as it is difficult to predict how it would impact the results presented in Chapter 5, as it is also affected by the distortion associated with the peripheries.

Considering the overwhelming accuracy achieved by the IVSN, as asserted by its proponents [29], it would be interesting to compare the accuracy of the proposed semantic-based model considered in this work [3], with an adapted version of zero-shot IVSN [29] model. To promote a proper comparison, this adapted version would consider a foveal visual system [5], allowing for an iterative search for any given stimuli, extracting top-down features along the way, while applying an IOR mechanism. Also, considering the evaluation metrics adopted by the proponents of the IVSN, a comparison between the search paths and patterns that result from the application of the different models could also be considered, together with the ones obtained with human subjects for the same experiments. Moreover, the distance between each saccade [25, 29] could also be assessed, to understand whether these models [3, 15, 29] resemble a performance that is somewhat similar to the human performance, leading the agent to act naturally.

The consideration made in Section 3.4.2, regarding the importance of considering the semantic information available in the cells that surround a potential fixation point, does not imply at all that the semantic information collected in those regions should be neglected or overlooked. It is possible that, by incorporating information available in adjacent cells, both the accuracy and precision of the active perception model may increase. Nevertheless, to achieve spatial precision, weights should be attributed to the information contained in surrounding cells, according to their distance to the focal point. For future

consideration, it would be certainly interesting to investigate how the integration of information available in adjacent cells could influence the performance of the active perception model. Such a technique would get more out of the map, taking full advantage of the spatial properties of the cells and the relationships between them, at the cost of increasing the overhead on the visual search algorithm.

Furthermore, as suggested by the work of Pal et al. [70], it would also be interesting to consider the possibility of integrating information related to hierarchical relationships between objects and classes, to facilitate the scene's navigation. In future work, hierarchical relations between classes could be integrated, for instance, in an active search metric, to quickly locate some class of interest in the scene.

Finally, it would be interesting to adapt the proposed methodology, to perform the same humanoid task, yet with a real mobile agent (e.g. a humanoid robot) that is able to perform body movements [25], moving through the environment and dynamically changing the range and direction of the field of view.

Bibliography

[1] J. Redmon and A. Farhadi, "Yolov3: An incremental improvement," University of Washington, Tech. Rep., 2018.

[2] J. Gu, Z. Wang, J. Kuen, L. Ma, A. Shahroudy, B. Shuai, T. Liu, X. Wang, G. Wang, J. Cai, and T. Chen, "Recent advances in convolutional neural networks," *Pattern Recognition*, vol. 77, pp. 354–377, 2018. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0031320317304120

[3] A. Dias, L. Simões, P. Moreno, and A. Bernardino, "Active gaze control for foveal scene exploration," in *2022 IEEE International Conference on Development and Learning (ICDL)*, Sep. 2022, pp. 115–120.

[4] R. Bajcsy, Y. Aloimonos, and J. K. Tsotsos, "Revisiting active perception," *Autonomous Robot*, vol. 42, pp. 177–196, 2018. [Online]. Available: https://arxiv.org/abs/1603.02729

[5] E. E. M. Stewart, M. Valsecchi, and A. C. Schütz, "A review of interactions between peripheral and foveal vision," *Journal of Vision*, vol. 20, no. 12, pp. 2–2, 11 2020. [Online]. Available: https://doi.org/10.1167/jov.20.12.2

[6] B. A. Wandell, *Foundations of vision.* Sinauer Associates, 1995.

[7] D. Pamplona and A. Bernardino, "Smooth foveal vision with gaussian receptive fields," in *2009 9th IEEE-RAS International Conference on Humanoid Robots*, 2009, pp. 223–229.

[8] A. Almeida, R. Figueiredo, A. Bernardino, and J. Santos-Victor, *Deep Networks for Human Visual Attention: A Hybrid Model Using Foveal Vision*. Springer, 01 2018, pp. 117–128.

[9] L. Itti, C. Koch, and E. Niebur, "A model of saliency-based visual attention for rapid scene analysis," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 20, no. 11, pp. 1254–1259, 1998.

[10] J. Redmon, S. Divvala, R. Girshick, and A. Farhadi, “You only look once: Unified, real-time object detection,” in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2016.

[11] P. Jiang, D. Ergu, F. Liu, Y. Cai, and B. Ma, “A review of yolo algorithm developments,” *Procedia Computer Science*, vol. 199, pp. 1066–1073, 2022, the 8th International Conference on Information Technology and Quantitative Management (ITQM 2020 & 2021): Developing Global Digital Economy after COVID-19. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S1877050922001363

[12] C.-Y. Wang, A. Bochkovskiy, and H.-Y. M. Liao, “YOLOv7: Trainable bag-of-freebies sets new state-of-the-art for real-time object detectors,” Institute of Information Science, Academia Sinica, Taiwan, Tech. Rep., 2022.

[13] A. Bochkovskiy, C.-Y. Wang, and H.-Y. M. Liao, “Yolov4: Optimal speed and accuracy of object detection,” 2020.

[14] H. Rezatofighi, N. Tsoi, J. Gwak, A. Sadeghian, I. Reid, and S. Savarese, “Generalized intersection over union: A metric and a loss for bounding box regression,” in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2019, pp. 658–666.

[15] S. Frintrop, T. Werner, and G. M. García, “Traditional saliency reloaded: A good old model in new shape,” in *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2015, pp. 82–90.

[16] T.-Y. Lin, M. Maire, S. Belongie, J. Hays, P. Perona, D. Ramanan, P. Dollár, and C. L. Zitnick, “Microsoft coco: Common objects in context,” in *Computer Vision – ECCV 2014*, D. Fleet, T. Pajdla, B. Schiele, and T. Tuytelaars, Eds. Cham: Springer International Publishing, 2014, pp. 740–755.

[17] L. K. Chan and W. G. Hayward, “Visual search,” *WIREs Cognitive Science*, vol. 4, no. 4, pp. 415–429, 2013. [Online]. Available: https://wires.onlinelibrary.wiley.com/doi/abs/10.1002/wcs.1235

[18] Z.-Q. Zhao, P. Zheng, S.-T. Xu, and X. Wu, “Object detection with deep learning: A review,” *IEEE Transactions on Neural Networks and Learning Systems*, vol. 30, no. 11, pp. 3212–3232, 2019.

[19] R. P. de Figueiredo and A. Bernardino, “An overview of space-variant and active vision mechanisms for resource-constrained human inspired robotic vision,” *Autonomous Robots*, Jun 2023. [Online]. Available: https://doi.org/10.1007/s10514-023-10107-7

[20] R. Murakami, S. Shimizu, T. Yamazaki, and N. Hasebe, “Saliency map for wide angle fovea vision sensor,” in *IECON 2018 - 44th Annual Conference of the IEEE Industrial Electronics Society*, 2018, pp. 5481–5486.

[21] E. L. Schwartz, “Computational anatomy and functional architecture of striate cortex: a spatial mapping approach to perceptual coding,” *Vision Research*, vol. 20, no. 8, pp. 645–669, 1980.

[22] 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)*, 2018, pp. 101–106.

[23] W. Liu, D. Anguelov, D. Erhan, C. Szegedy, S. Reed, C.-Y. Fu, and A. C. Berg, “Ssd: Single shot multibox detector,” in *Computer Vision – ECCV 2016*, B. Leibe, J. Matas, N. Sebe, and M. Welling, Eds. Cham: Springer International Publishing, 2016, pp. 21–37.

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

[25] R. Druon, Y. Yoshiyasu, A. Kanezaki, and A. Watt, “Visual object search by learning spatial context,” *IEEE Robotics and Automation Letters*, vol. 5, no. 2, pp. 1279–1286, 2020.

[26] E. Sommerlade and I. Reid, “Information-theoretic active scene exploration,” in *2008 IEEE Conference on Computer Vision and Pattern Recognition*, 2008, pp. 1–7.

[27] P. Cavanagh, “Visual cognition,” *Vision Research*, vol. 51, no. 13, pp. 1538–1551, 2011, vision Research 50th Anniversary Issue: Part 2. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0042698911000381

[28] R. Burt, N. N. Thigpen, A. Keil, and J. C. Principe, “Unsupervised foveal vision neural architecture with top-down attention,” *Neural Networks*, vol. 141, pp. 145–159, 2021. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0893608021000836

[29] M. Zhang, J. Feng, K. T. Ma, J. H. Lim, Q. Zhao, and G. Kreiman, “Finding any waldo with zero-shot invariant and efficient visual search,” *Nature Communications*, vol. 9, no. 1, sep 2018. [Online]. Available: https://doi.org/10.10382Fs41467-018-06217-x

[30] D. M. Levi, “Crowding—an essential bottleneck for object recognition: A mini-review,” *Vision Research*, vol. 48, no. 5, pp. 635–654, 2008.

[31] V. Javier Traver and A. Bernardino, “A review of log-polar imaging for visual perception in robotics,” *Robotics and Autonomous Systems*, vol. 58, no. 4, pp. 378–398, 2010. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0921889009001687

[32] P. Ozimek, N. Hristozova, L. Balog, and J. P. Siebert, “A space-variant visual pathway model for data efficient deep learning,” *Frontiers in Cellular Neuroscience*, vol. 13, 2019. [Online]. Available: https://www.frontiersin.org/articles/10.3389/fncel.2019.00036

[33] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, "Imagenet: A large-scale hierarchical image database," in *2009 IEEE Conference on Computer Vision and Pattern Recognition*, 2009, pp. 248–255.

[34] H. Lukanov, P. König, and G. Pipa, "Biologically inspired deep learning model for efficient foveal-peripheral vision," *Frontiers in Computational Neuroscience*, vol. 15, p. 746204, 2021.

[35] P. J. Burt and E. H. Adelson, "The laplacian pyramid as a compact image code," in *Readings in computer vision*. Elsevier, 1987, pp. 671–679.

[36] J. Duncan and G. W. Humphreys, "Visual search and stimulus similarity," *Psychological Review*, vol. 96, no. 3, p. 433, 1989.

[37] S. Frintrop and J. Hertzberg, *VOCUS: A Visual Attention System for Object Detection and Goal-Directed Search*. Springer, 2006, vol. 3899, pp. 57–127.

[38] S. Frintrop, *Computer Analysis of Human Behavior*. London: Springer London, 2011, ch. Computational Visual Attention, pp. 69–101. [Online]. Available: https://doi.org/10.1007/978-0-85729-994-9_4

[39] C. Chen, X. Zhang, Y. Wang, T. Zhou, and F. Fang, "Neural activities in v1 create the bottom-up saliency map of natural scenes," *Experimental Brain Research*, vol. 234, 06 2016.

[40] P. F. Felzenszwalb, R. B. Girshick, D. McAllester, and D. Ramanan, "Object detection with discriminatively trained part-based models," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 32, no. 9, pp. 1627–1645, Sep. 2010.

[41] A. Harakeh, M. Smart, and S. L. Waslander, "Bayesod: A bayesian approach for uncertainty estimation in deep object detectors," in *2020 IEEE International Conference on Robotics and Automation (ICRA)*, 2020, pp. 87–93.

[42] R. Padilla, S. L. Netto, and E. A. B. da Silva, "A survey on performance metrics for object-detection algorithms," in *2020 International Conference on Systems, Signals and Image Processing (IWSSIP)*, 2020, pp. 237–242.

[43] M. Stoycheva, "Uncertainty estimation in deep neural object detectors for autonomous driving," Master's thesis, KTH Royal Institute of Technology, School of Electrical Engineering and Computer Science, 2020.

[44] Y. Sasaki *et al.*, "The truth of the f-measure," *Teach tutor mater*, vol. 1, no. 5, pp. 1–5, 2007.

[45] D. C. Montgomery and G. C. Runger, *Applied statistics and probability for engineers*. John wiley & sons, 2010.

[46] C. M. Bishop, *Pattern Recognition and Machine Learning (Information Science and Statistics)*. Berlin, Heidelberg: Springer-Verlag, 2006.

[47] T. Silva Filho, H. Song, M. Perello-Nieto, R. Santos-Rodriguez, M. Kull, and P. Flach, "Classifier calibration: a survey on how to assess and improve predicted class probabilities," *Machine Learning*, vol. 112, no. 9, pp. 3211–3260, Sep 2023. [Online]. Available: https://doi.org/10.1007/s10994-023-06336-7

[48] M. Xie, S. Li, R. Zhang, and C. H. Liu, "Dirichlet-based uncertainty calibration for active domain adaptation," 2023. [Online]. Available: https://openreview.net/forum?id=4WM4cy42B81

[49] Y. Gall, "Uncertainty in deep learning," Master's thesis, University of Cambridge, September 2016.

[50] N. L. Johnson, S. Kotz, and N. Balakrishnan, *Discrete multivariate distributions*. Wiley New York, 1997, vol. 165.

[51] T. Minka, "Estimating a Dirichlet distribution," Massachusetts Institute of Technology, Tech. Rep., 2000, revised in 2003, 2009, 2012.

[52] M. Abdar, F. Pourpanah, S. Hussain, D. Rezazadegan, L. Liu, M. Ghavamzadeh, P. Fieguth, X. Cao, A. Khosravi, U. R. Acharya, V. Makarenkov, and S. Nahavandi, "A review of uncertainty quantification in deep learning: Techniques, applications and challenges," *Information Fusion*, vol. 76, pp. 243–297, dec 2021.

[53] E. Hüllermeier and W. Waegeman, "Aleatoric and epistemic uncertainty in machine learning: an introduction to concepts and methods," *Machine Learning*, vol. 110, pp. 457–506, 2021.

[54] A. Kendall and Y. Gal, "What uncertainties do we need in bayesian deep learning for computer vision?" in *Proceedings of the 31st International Conference on Neural Information Processing Systems*, ser. NIPS'17. Red Hook, NY, USA: Curran Associates Inc., 2017, p. 5580–5590.

[55] M. I. Coco, A. Nuthmann, and O. Dimigen, "Fixation-related Brain Potentials during Semantic Integration of Object–Scene Information," *Journal of Cognitive Neuroscience*, vol. 32, no. 4, pp. 571–589, 04 2020. [Online]. Available: https://doi.org/10.1162/jocn_a_01504

[56] F. Tonini, N. Dall'Asen, C. Beyan, and E. Ricci, "Object-aware gaze target detection," in *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, October 2023, pp. 21 860–21 869.

[57] 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*, vol. 42, pp. 459–476, 2018. [Online]. Available: https://link.springer.com/article/10.1007/s10514-017-9617-1

[58] 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)*. IEEE Press, 2017, p. 1427–1434. [Online]. Available: https://doi.org/10.1109/IROS.2017.8205944

[59] J. Wolfe and T. Horowitz, “What attributes guide the deployment of visual attention and how do they do it?” *Nature Reviews Neuroscience*, vol. 5, p. 495–501, 2004.

[60] J. L. Schönberger, M. Pollefeys, A. Geiger, and T. Sattler, “Semantic visual localization,” in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018, pp. 6896–6906.

[61] M. Kiefer, U. Ansorge, J.-D. Haynes, F. Hamker, U. Mattler, R. Verleger, and M. Niedeggen, “Neurocognitive mechanisms of conscious and unconscious visual perception: From a plethora of phenomena to general principles,” *Advances in Cognitive Psychology*, vol. 7, p. 55, 2011.

[62] L. M. Kaplan, S. Chakraborty, and C. Bisdikian, “Fusion of classifiers: A subjective logic perspective,” in *2012 IEEE Aerospace Conference*, 2012, pp. 1–13.

[63] L. D. Simões and A. J. M. Bernardino, “Active perception: Scene exploration using foveal vision,” September 2021. [Online]. Available: https://fenix.tecnico.ulisboa.pt/downloadFile/844820067126711/IST_UL_MSc_Thesis__Copy_(7).pdf

[64] R. De Maesschalck, D. Jouan-Rimbaud, and D. Massart, “The mahalanobis distance,” *Chemometrics and Intelligent Laboratory Systems*, vol. 50, no. 1, pp. 1–18, 2000. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0169743999000477

[65] D. Ruta and B. Gabrys, “An overview of classifier fusion methods,” *Computing and Information systems*, vol. 7, no. 1, pp. 1–10, 2000.

[66] S.-H. Cha, “Comprehensive survey on distance/similarity measures between probability density functions,” *City*, vol. 1, no. 2, p. 1, 2007.

[67] B. Kurt, “Kullback-leibler divergence between two dirichlet (and beta) distributions,” https://bariskurt.com/kullback-leibler-divergence-between-two-dirichlet-and-beta-distributions/, 2013.

[68] M. B. Mirza, R. A. Adams, C. Mathys, and K. J. Friston, “Human visual exploration reduces uncertainty about the sensed world,” *PLOS ONE*, vol. 13, no. 1, pp. 1–20, 01 2018. [Online]. Available: https://doi.org/10.1371/journal.pone.0190429

[69] K. Yun, Y. Peng, D. Samaras, G. J. Zelinsky, and T. L. Berg, “Exploring the role of gaze behavior and object detection in scene understanding,” *Frontiers in psychology*, vol. 4, p. 917, 2013.

[70] A. Pal, Y. Qiu, and H. Christensen, "Learning hierarchical relationships for object-goal navigation," in *Conference on Robot Learning*. PMLR, 2021, pp. 517–528.

[71] M. Spiegel, S. Lipschutz, and J. Liu, *Schaum's Outline of Mathematical Handbook of Formulas and Tables: 2,400 Formulas + Tables*. McGraw Hill Professional, 2013.

A

Dirichlet Distribution Estimation

This appendix presents and explains the complete formulation of the *Dirichlet fast Fit* algorithm, proposed by Minka [51]. This algorithm fits the parameters $\boldsymbol{\alpha}$ of a Dirichlet distribution to a set of training data points $\mathcal{D}$ in an iterative fashion, through a maximum likelihood estimation of (2.12).

The iterative approach, suggested by Minka [51], considers an alternate optimization between the precision, $\nu = \sum_{k=1}^{K} \alpha_k$, and the mean values, $\boldsymbol{m} = \left(\frac{\alpha_1}{\nu}, \ldots, \frac{\alpha_K}{\nu}\right)$, by fixing one parameter and only optimizing the other, during each iteration, in order to simplify and speed up the training process. Therefore, the $\boldsymbol{\alpha}$ parameters can be reparametrized as $\alpha_k = \nu m_k$ and substituted in (2.12), in order to extract the isolated likelihood that considers the precision ν alone [51], which can be described by the relation

$$p(\mathcal{D}|\nu) \propto \left(\frac{\exp\left(\frac{\nu}{N} \sum_{k=1}^{K} \sum_{i=1}^{N} m_k \log p_{i,k}\right) \Gamma(\nu)}{\prod_{k=1}^{k} \Gamma(\nu m_k)} \right)^N . \tag{A.1}$$

As the next step in the optimization process of the precision ν, the derivatives of (A.1) are defined as

$$\frac{d}{d\nu} \log p(\mathcal{D}|\nu) = N \left[\Psi(\nu) - \sum_k m_k \left(\Psi(\nu m_k) + \frac{1}{N} \sum_{i=1}^{N} \log p_{i,k} \right) \right] \tag{A.2}$$

$$\frac{d^2}{d\nu^2} \log p(\mathcal{D}|\nu) = N \left(\Psi'(\nu) - \sum_k m_k^2 \Psi'(\nu m_k) \right) \tag{A.3}$$

where $\Psi(\nu) = \frac{d}{d\nu} \log \Gamma(\nu)$ is known as the *digamma* function [46], and $\Psi'(\nu) = \frac{d}{d\nu} \Psi(\nu)$ corresponds to the respective derivative. Then, Minka [51] considers a generalized Newton-Raphson iteration [46], leading to a new update rule that converges quite rapidly. This generalized update rule is expressed as

$$\frac{1}{\nu^{\text{new}}} = \frac{1}{\nu} + \frac{1}{\nu^2} \left(\frac{d}{d\nu} \log p(\mathcal{D}|\nu) \right) \left(\frac{d^2}{d\nu^2} \log p(\mathcal{D}|\nu) \right)^{-1} . \tag{A.4}$$

To initialize the precision ν, Minka uses Stirling's approximation [71] to the $\Gamma(z)$ function (2.11), namely

$$\frac{\Gamma(\nu)}{\prod_{k=1}^{K} \Gamma(\nu m_k)} \exp\left(\frac{\nu}{N} \sum_{k=1}^{K} \sum_{i=1}^{N} m_k \log p_{i,k} \right) \approx \left(\frac{\nu}{2\pi} \right)^{\frac{K-1}{2}} \prod_{k=1}^{K} \sqrt{m_k} \exp\left(\frac{\nu}{N} \sum_{k=1}^{K} \sum_{i=1}^{N} m_k \log \frac{p_{i,k}}{m_k} \right) \tag{A.5}$$

Considering this approximation of the precision (A.5), one can then extract its initial value $\hat{\nu}$, given by

$$\hat{\nu} \approx -N \frac{(K-1)/2}{\sum_{k=1}^{K} \sum_{i=1}^{N} m_k \log \frac{p_{i,k}}{m_k}} . \tag{A.6}$$

As previously stated, ν and $\boldsymbol{m}$ are roughly decoupled in the maximum-likelihood objective, allowing for an alternate optimization of these parameters. Hence, consider now that the precision ν is fixed [51], to estimate the mean $\boldsymbol{m}$, also considering a parameterization, as for ν. The individual likelihood for m is

$$p(\mathcal{D}|\boldsymbol{m}) \propto \left(\prod_{k=1}^{K} \frac{\exp\left(\nu m_k \sum_{i=1}^{N} \log p_{i,k}\right)}{\Gamma(\nu m_k)} \right)^N . \tag{A.7}$$

Let us now consider the parameterization of the likelihood function (A.7). In order to obtain the gradient

of (A.7) one must consider using the unconstrained vector $\boldsymbol{z}$, leading to a redefinition of each mean

$$m_k = \frac{z_k}{\sum_{j=1}^{K} z_j}. \tag{A.8}$$

Accounting for the new mean parameterization (A.8), the log-likelihood for m alone can be written as

$$\log p(\mathcal{D}|\boldsymbol{m}) = N \sum_{k=1}^{K} \left[\frac{1}{N} \frac{z_k}{\sum_{j=1}^{K} z_j} \sum_{i=1}^{N} \log p_{i,k} - \log \Gamma \left(\nu \frac{z_k}{\sum_{j=1}^{K} z_j} \right) \right]. \tag{A.9}$$

To find the maximum, the gradient of the log-likelihood (A.9) is then computed, through the expression

$$\frac{d}{dz_k} \log p(\mathcal{D}|\boldsymbol{m}) = \frac{\nu N}{\sum_{j=1}^{K} z_j} \left[\frac{1}{N} \sum_{i=1}^{N} \log p_{i,k} - \Psi(\nu m_k) - \sum_{j=1}^{K} m_j \left(\frac{1}{N} \sum_{i=1}^{N} \log p_{i,j} - \Psi(\nu m_j) \right) \right]. \tag{A.10}$$

With (A.10), one is now in a position to use the Maximum Likelihood Estimation (MLE) to find the new mean values ($\boldsymbol{m}^{\text{new}}$). The MLE can be computed through a fixed-point iteration by solving the equation

$$\frac{d \log p(\mathcal{D} \mid \boldsymbol{m})}{dz_k} = 0 \Leftrightarrow \Psi(\alpha_k) = \frac{1}{N} \sum_{i=1}^{N} \log p_{i,k} - \sum_{j=1}^{N} m_j^{\text{old}} \left[\frac{1}{N} \sum_{i=1}^{N} \log p_{i,j} - \Psi\left(\nu m_j^{\text{old}} \right) \right], \tag{A.11}$$

remembering that $\alpha_k = \nu m_k$. Finally, the new mean value can be established, after solving (A.11), as

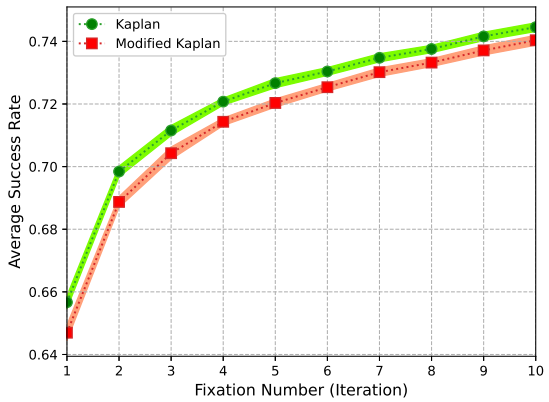
$$m_k^{\text{new}} = \frac{\alpha_k}{\sum_k \alpha_k}. \tag{A.12}$$

The algorithm is terminated when the Euclidean distance between $\boldsymbol{m}^{\text{new}}$ and $\boldsymbol{m}^{\text{old}}$ sinks bellow a pre-defined tolerance value ξ, after each estimation of both mean and precision in consecutive iterations.

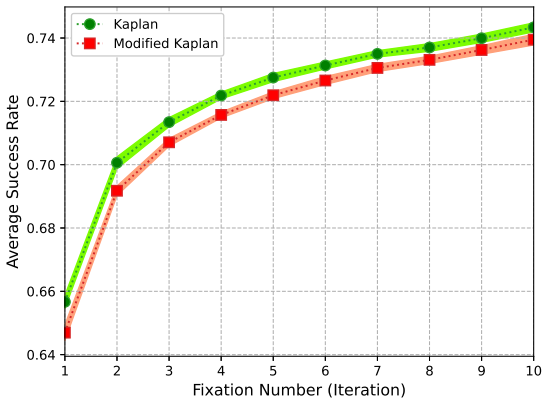
B

Scene Exploration Results

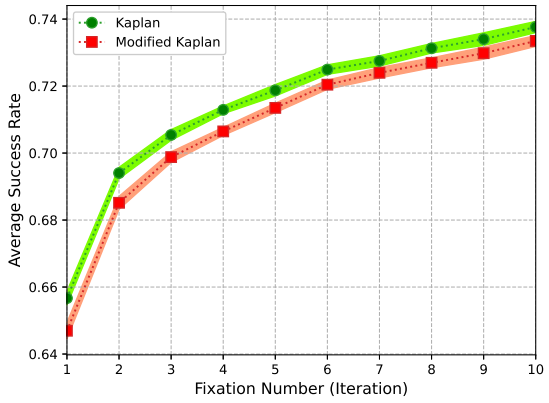
In this appendix are presented the complementary results of Section 5.2.1, proving that the Kaplan map tends to better explain the scene's semantic content when taking into consideration the full methodological pipeline represented in Figure 3.1 and meticulously explained in Chapter 3.



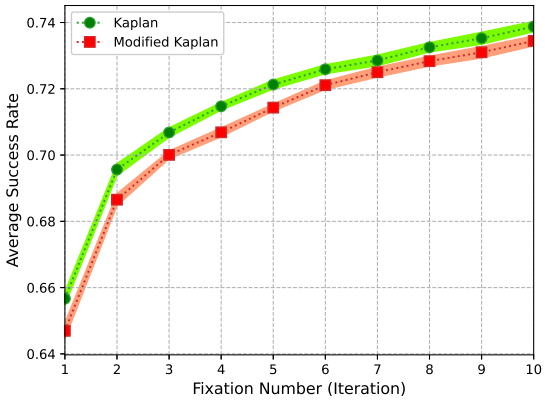
(a) Maximize KL Divergence w/ Kaplan



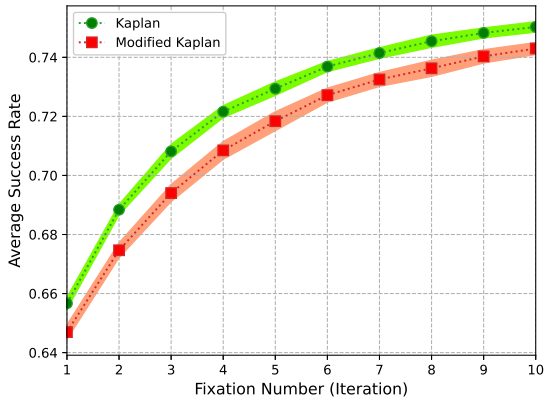
(b) Maximize KL Divergence w/ Modified Kaplan



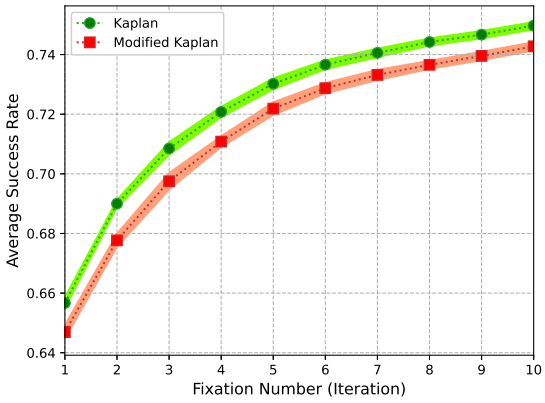
(c) Minimize Entropy w/ Kaplan



(d) Minimize Entropy w/ Modified Kaplan



(e) Maximize Two-Peak Difference w/ Kaplan



(f) Maximize Two-Peak Difference w/ Mod. Kaplan

Figure B.1: Extension of the results presented in Figure 5.1, with different active perception methods, namely the maximization of the KL divergence (a) (b), the negentropy (c)(d), and the two-peak difference (e) (f), using either the semantic information gathered either in the Kaplan map or the Modified Kaplan map.

C

Visual Search Results

In this appendix are presented the first four iterations obtained while completing a visual search [17] experiment, using a predictive approach (3.25). This example has been introduced in Section 5.3.4, where it asserted that ”book” is the class being searched in the scene. Hence, the results presented in Figure C.1 complement the remarks made in Section 5.3.4 regarding the effectiveness of the proposed methodology when it comes to rapidly finding target objects in a visual search context.

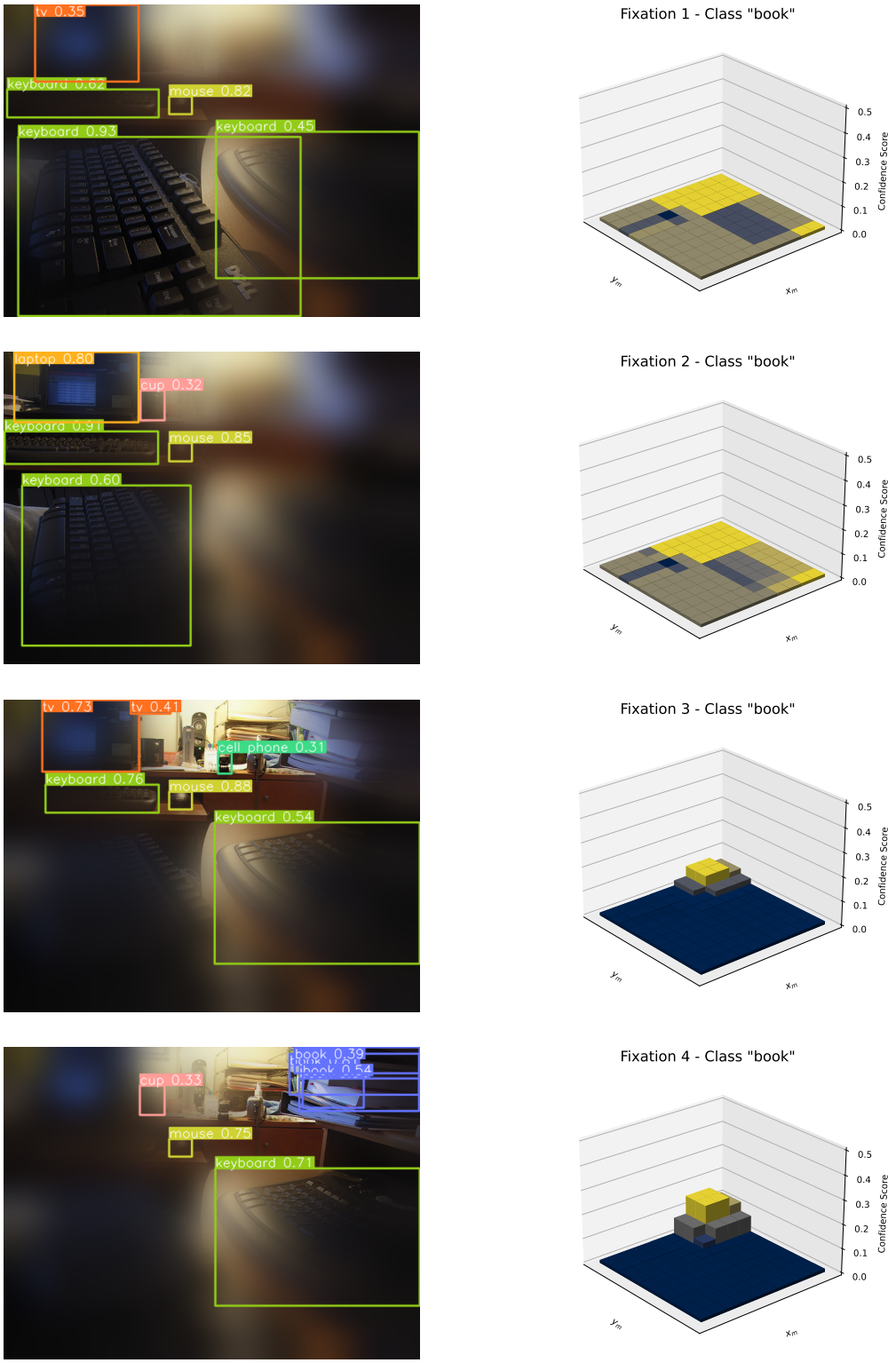


Figure C.1: Example of a visual search [17] experiment, where the class ”book” is defined as the target. Here are presented the first four iterations of the search algorithm (3.25), where the confidence scores of the target class consistently increase in the region where the ground-truth objects are effectively located.