Active Gaze Control for Foveal Scene Exploration

Alexandre Dias, Luís Simões, Plinio Moreno, Alexandre Bernardino
Instituto Superior Técnico
Lisbon, Portugal
{alexandre.f.dias, luis.d.simoes, alexandre.bernardino}@tecnico.ulisboa.pt, plinio@isr.tecnico.ulisboa.pt

Abstract—Active perception and foveal vision are the foundations of the human visual system. While foveal vision reduces the amount of information to process during a gaze fixation, active perception will change the gaze direction to the most promising parts of the visual field. We propose a methodology to emulate how humans and robots with foveal cameras would explore a scene, identifying the objects present in their surroundings with at least number of gaze shifts. Our approach is based on three key methods. First, we take an off-the-shelf deep object detector, pre-trained on a large dataset of regular images, and calibrate the classification outputs to the case of foveated images. Second, a body-centered semantic map, encoding the objects classifications and corresponding uncertainties, is sequentially updated with the calibrated detections, considering several data fusion techniques. Third, the next best gaze fixation point is determined based on information-theoretic metrics that aim at minimizing the overall expected uncertainty of the semantic map. When compared to the random selection of next gaze shifts, the proposed method achieves an increase in detection F1-score of 2-3 percentage points for the same number of gaze shifts and reduces to one third the number of required gaze shifts to attain similar performance.

Index Terms—active perception, foveal vision, object detection, active object search, data fusion

I. INTRODUCTION

The fovea is the area of the visual retina with the highest concentration of photoreceptors. In humans, the fovea comprises less than 1% of retinal size, but takes up over 50% of the visual cortex [1]. In artificial systems, a foveated image is an image having an area of high resolution (typically the center), where fine details of objects can be observed, while the remaining of the image (periphery) has low resolution and objects look blurry. An illustration of an artificial foveal image is shown in Fig. 1.

Central or foveal vision is an indispensable feature of the human eye, allowing activities that require high visual acuity such as reading, sports and driving to be carried out. It has been implemented with success in robots to perform visual tasks such as object tracking and depth perception with low computational resources [2]. Visual search tasks require nonetheless the involvement of peripheral vision, from which it has proved challenging for image processing and pattern recognition methods to extract relevant information.

Although there has been a large amount of research and developments on visual attention and visual search models [3]–[6], there is still a long way to go, specially regarding the modeling of the mechanisms that help the decision of where to shift the gaze to. Some approaches consider interesting points to look at as points that are salient in some image dimension (e.g. color) or have high uncertainty in other dimensions (e.g. depth) [7], [8], but lack on the semantic representation of the observed objects. Other approaches make use of recent deep object detectors, that carry semantic information about the classes of objects, to detect objects of interest in the periphery of the visual field and perform visual search tasks [5], [6]. However, a major challenge in visual search with foveal images is the fact that object detectors and saliency models have not been designed to work on low resolution images, thus lagging behind in performance in the detection of objects in the peripheral regions of a foveal image.

In this paper, we propose a methodology that allows exploring a foveal scene while minimizing the required number of gaze fixations in order to identify the objects present in the scene. Scene exploration is achieved by using a readily available deep object detector and by implementing three key elements: a Foveal Observation Model, which learns the response of the Object Detector to foveated images; Data Fusion techniques, which combine the information of subsequent detections in order to update the knowledge about the scene; Active Perception techniques, which determine the best location in the scene to where the gaze shall be shifted. To the extent of our knowledge, this is the first attempt at combining Object Detection and Active Perception to perform scene exploration with foveal vision.

The paper is structured as follows: in Section II, the proposed methodology and its key elements are justified and presented in detail; in Section III, the experimental setup is described and the results of the experiments are presented and discussed; in Section IV, we conclude regarding the validity of the proposed methods.
II. PROPOSED METHODOLOGY

A. Overview

The main goal of this work is to implement a methodology optimizing the exploration of a scene using foveal vision, to gather as much information as possible about the objects present in the scene in the least number of gaze shifts.

We assume the observing agent to be stationary, analogous to a human not moving the head nor the body, yet to be able to, without any uncertainty, i) shift its gaze to certain parts of the visual field and ii) sense the current gaze direction. The transformation from world to retinal coordinates is defined as a simple translation. The agent’s gaze being fixed at world coordinates \((x_t, y_t)\), an observation at time step \(t\) in world coordinates \((x, y)\) corresponds to an observation in local image coordinates \((u_t, v_t) = (x - x_t, y - y_t)\), and the distance \(d_t\) to the focal point is given by \(d_t = \| (u_t, v_t) \|\) — the subscript of \(d_t\) will be omitted for simplicity, unless required to avoid ambiguity. In Fig. 2, a schematic illustration of the spatial scene representation considered is shown.

After the fixation of the gaze at time step \(t\), the image is foveated. The foveation method takes into account the current focal point (center of the fovea) and foveates the full image such that the foveal region keeps the spectral content of the original image and the regions in the periphery are low-pass filters with decreasing bandwidth as the distance from the fovea increases. The used foveation method is described in [5], [6] and code is available online\(^1\).

The foveated image then serves as input to an object detector, which computes a set of \(L_t\) detections, denoted as \(\mathcal{I}_t = \{I_{t}^{l_i}\}_{i = 1, \ldots, L_t}\), where \(I_{t}^{l_i} = (B_{t}^{l_i}, S_{t}^{l_i})\) is a 2-tuple comprised of the location and size (bounding box) of the \(l_i\)th detection, \(B_{t}^{l_i}\), and the respective class confidence scores, \(S_{t}^{l_i} = (s_{t,0}^{l_i}, s_{t,1}^{l_i}, \ldots, s_{t,K}^{l_i})\) \(\in \Delta^K\). The set of possible object classes is \(\mathcal{C} = \{0, 1, \ldots, K\}\), where \(K\) is the number of object classes and 0 denotes the background class, i.e., a class representing the nonexistence of an object. In Fig. 3, object detections in a foveated image are depicted: the names and percentages correspond to the class of each detection having the highest confidence score.

Nevertheless, off-the-shelf object detectors do not provide classification scores appropriate for foveated images. As they are trained on regular images, for objects undistorted by foveal vision (objects in the foveal region) the detector output is expected to be similar to that of a regular image; however, for objects in the periphery of the fovea, the detector is expected to output classification scores with increased uncertainty (increased entropy) due to the additional class ambiguity in the blurred part of the visual field. Thus, a Foveal Observation Model learns how the class confidence scores of the detections vary with the position relative to the focal point, allowing to calibrate such scores and providing uncertainty measures of the detections. This approach sidesteps the retraining of the object detector’s neural network, which can be excessively time consuming.

On the other hand, a coarse detection in a foveated image corresponding to the peripheral view of an object will inform the agent of the potential semantic classes that may be in that location and should prompt the gaze to be shifted to that location if the information about the object is still scant, i.e., so far there has been no gaze fixation close to the object. A semantic map would then be improved by a gaze fixation closer to such spacial location. Then, to perform scene exploration efficiently, the detector output, after calibration by the observation model, is merged with previous detection information to update a body-centered semantic map with the knowledge about the scene.

Active Perception can then exploit this updated knowledge to determine the next best viewpoint, i.e., the focal point producing new observations expected to decrease the overall uncertainty about the scene. The gaze is then shifted towards this point, and the image is again foveated to reflect this new choice of focal point. This process of image foveation, object detection, map information update and next best gaze fixation determination is then repeated until a suitable stopping criterion is met.

---

\(^1\)github.com/vislab-tecnico-lisboa/laplacian-foveation

![Fig. 2. Scene representation: world coordinates \(x, y\) and local image coordinates \(u, v\); the rectangles correspond to the objects present and the center of the dashed circles is the focal point \((x_t, y_t)\).](image1)

![Fig. 3. Object detections in a foveated image. The names and percentages indicate the class with the highest confidence score of the respective detection (bounding box).](image2)
B. Foveal Observation Model

The Foveal Observation Model should account for the increased uncertainty of each detection due to the foveation process. The detector output scores, \( S^d_t \), when normalized to unit sum, can be seen as a realization of a Dirichlet distribution. Thus, for a given labeled dataset, the parameters of the Dirichlet distributions that characterize the object detector response to each object class \( k \in C \) at each distance \( d \) to the fovea, \( \alpha_{k,d} = (\alpha_{k,d,1}, \alpha_{k,d,2}, \ldots, \alpha_{k,d,K}) \), can be estimated. The Foveal Observation Model thus consists of sets of Dirichlet distribution parameters — one set of parameter values per object class \( k \) and per distance level \( d \) (7 different distance levels were considered), and it is trained to learn the probability distribution \( p(S|C=k,d) \).

The normalized scores vector \( S^d_t \) corresponding to each new detection \( I^d_t \) being modeled as realizations of the previously trained Dirichlet distributions: \( S \sim \text{Dir}(\alpha_{k,d}) \), for each class \( k = 1, ..., K \), a new set of calibrated classification scores \( S' \) can be computed that more accurately reflect the probabilities of detections in the foveal images:

\[
S' = (s'_1, \ldots, s'_K) = \frac{1}{D} \left[ p(S^d_t|C = 1, d), \ldots, p(S^d_t|C = K, d) \right],
\]

where \( D = \sum_{k=1}^{K} p(S^d_t|C = k, d) \) is a normalization factor such that \( \sum_{k=1}^{K} s'_k = 1 \). The calibrated classification scores \( S' \) are expected to have less entropy than the original \( S \).

C. Fusion Model

The fusion problem consists in, given a set of classification scores for a single pattern, calculating an overall classification score \( p \) or estimating a distribution of classification scores. As the agent explores the scene by shifting the gaze to different parts of the visual field, the information of subsequent detections is collected. This information is used to update a map \( M \). The \((K + 1)\)-channel map \( M_t(x,y) \) is defined as:

\[
M_t(x,y) = p_t^{x,y} = (p_t^{x,y}_1, p_t^{x,y}_2, \ldots, p_t^{x,y}_K) \in \Delta^K ,
\]

where \( p_t^{x,y}_k = P(C^{x,y} = k|I_{1:t}, x_{1:t}, y_{1:t}) \) specifies the class posterior probability of class \( k \in C \). Thus, \( p_t^{x,y} \) constitutes the estimate, at time step \( t \), of the class posterior probabilities of an object located at coordinates \((x,y)\).

Given a sequence of classifier scores \( S_{1:t} \) for time steps \( 1 \) up to \( t \), world coordinates \((x,y)\) and distances \( d_{1:t} \) to the fovea, a simple product rule (Naïve Bayes) can be used to update the map information, according to

\[
M_{t,k}(x,y) = P(C^{x,y} = k|S_{1:t}, d_{1:t}) = \frac{1}{Z_t} \prod_{\tau=1}^{t} p(S_{\tau}|C^{x,y} = k) = \frac{1}{Z_t} p(S_{t}|C^{x,y} = k)M_{t-1,k}(x,y)
\]

where \( Z \) and \( Z_t \) are unit-sum normalizing constants.

A categorical distribution is unable to deal with the uncertainty in the estimated parameters \( p \) (variance), only with the first moment (expected value). Such uncertainty can, however, be captured by modeling the parameters \( p \) as realizations of a Dirichlet distribution, \( p \sim \text{Dir}(\beta) \), \( \beta = (\beta_0, \ldots, \beta_K) \). This leads to the definition of a map \( B_t(x,y) \) containing the estimates at time step \( t \) and for world coordinates \((x,y)\) of the parameters:

\[
B_t(x,y) = \beta_{t}^{x,y} = (\beta_{t,x,y}^{x,y}, \beta_{t,y}^{x,y}, \ldots, \beta_{t,K}^{x,y})
\]

These parameters are able to encode the uncertainty — through the second moment or variance — of the class posterior scores \( p_{t,k}^{x,y} \) and allow for practical sequential update rules. The class posterior probabilities of \( M_t(x,y) \) can be computed from the parameters of \( B_t(x,y) \) through

\[
M_{t,k}(x,y) = p_{t,k}^{x,y} = E_{p_{t,k}^{x,y}}[\text{Dir}(\beta_{t}^{x,y})] = \frac{\beta_{t,k}^{x,y}}{\sum_{j=1}^{K} \beta_{t,j}^{x,y}}.
\]

Several strategies for the sequential update of the map \( M_t \) exist. In [9], Kaplan et al. proposes several update rules for the fusion of classifier scores: the product rule (equivalent to Naïve Bayes presented), the sum rule, and a novel approach that we denote by Kaplan’s rule.

The application of the sum rule consists in simply adding each of the classifier scores to the parameter of the corresponding class \( k \):

\[
\beta_{t+1,k} = \beta_{t,k} + s_{t,k}.
\]

Differently from the product rule, in the sum rule the observations are summed instead of multiplied, providing robustness to outliers. On the other hand, the sum rule does not yield a posterior Dirichlet distribution.

In Kaplan’s rule [9], a Dirichlet distribution is fitted to the actual posterior through a moment matching approach. Nevertheless, using this rule with the raw output of the detector would ignore the knowledge about the influenced of the objects location relative to the center of the fovea (Foveal Observation...
Model). Thus, in order to assess a potential improvement in exploring the scene, a modified version of Kaplan’s rule can consider the calibrated scores $S_{t,i}$ of (1) instead of the original detector scores $S_i^t$:

$$
\beta_{t+1,k} = \frac{\beta_{t,k} \left( 1 + \frac{\sum_{j=1}^{K} \beta_{t,j} S_{t,j}^t}{\sum_{j=1}^{K} \beta_{t,j} S_{t,j}^t} \right)}{1 + \min_{j=1}^{K} \frac{\sum_{j=1}^{K} \beta_{t,j} S_{t,j}^t}{\sum_{j=1}^{K} \beta_{t,j} S_{t,j}^t}} 
$$

(6)

D. Active Perception Model

In order to perform scene exploration in the least number of gazes, and based on the updated knowledge of the map, an Active Perception method chooses the point $(x_{t+1}, y_{t+1})$ to which to shift the gaze in the next time step $t+1$, minimizing some measure of uncertainty $F$:

$$(x_{t+1}^*, y_{t+1}^*) = \arg\min_{x_{t+1}, y_{t+1}} F(x_{t+1}, y_{t+1}, I_{t+1}; x_{t}, y_{t}; t)$$

(7)

This requires predicting the information accumulated up to time step $t+1$, $I_{t+1}$, that would be obtained if the gaze direction were changed to coordinates $(x', y') = (x_{t+1}^*, y_{t+1}^*)$. As the knowledge about the scene at time $t$ is represented by the semantic map $B_t(x, y)$, we need to compute:

$$
B_{t+1}^{x',y'}(x, y) = \mathbb{E}[B_{t+1}(x, y) | B_t(x, y), x', y']
$$

(8)

To perform the computation of (8) we need to predict the detector classification scores for each world coordinate $(x, y)$, given that the gaze is shifted to $(x', y')$. Let $d' = |(x - x', y - y')|$. Then, $p(S^{x,y} | x', y', B_{t,k}^{x,y})$ can be written as:

$$
p(S^{x,y} | x', y', B_{t,k}^{x,y}) = \sum_{k=0}^{K} p(S^{x,y} | k, x', y', B_{t,k}^{x,y})
$$

$$
= \sum_{k=0}^{K} p(S^{x,y} | k, x', y', B_{t,k}^{x,y}) p(k | x', y', B_{t,k}^{x,y})
$$

$$
= \sum_{k=0}^{K} p(S^{x,y} | k, x', y', B_{t,k}^{x,y}) p(k | B_{t,k}^{x,y})
$$

(9)

where the last step used the fact that $S^{x,y}$ is conditionally independent of $B_{t,k}^{x,y}$ given the object class $k$, and that the object class $k$ at coordinates $(x, y)$ does not depend on the placement of the fovea at time $t+1$. Finally, since $p(k | B_{t,k}^{x,y}) = \beta_{t,k}/\sum_{i=0}^{K} \beta_{t,i}^{x,y}$:

$$
p(S^{x,y} | x', y', B_{t,k}^{x,y}) \propto \sum_{k=0}^{K} p(S^{x,y} | k, x', y') \beta_{t,k}^{x,y} / \sum_{i=0}^{K} \beta_{t,i}^{x,y}.
$$

Given the linearity of the expectation operator:

$$
\overline{S}^{x,y} = \mathbb{E}[S^{x,y} | x', y'] = \sum_{k=0}^{K} \frac{\alpha_{k,d}}{\sum_{i=0}^{K} \alpha_{i,d}} \beta_{t,k}^{x,y} / \sum_{i=0}^{K} \beta_{t,i}^{x,y}.
$$

(10)

A measure of uncertainty is associated to the class posterior distribution at each map coordinate, represented as $U(x, y)$. Several measures of uncertainty are considered: the Kullback-Leibler (KL) divergence or relative entropy, the entropy, and the difference between the scores of the two classes with highest confidence.

The KL divergence is used to measure the information gain between the current Dirichlet distribution of parameters $\beta$ and an initial Dirichlet uniform distribution ($\beta_{0,k} = 1, k = 0, \ldots, K$):

$$
U_t(x, y) = -D_{KL}(\text{Dir}(\beta_{t}^{x,y}) || \text{Dir}(\beta_0)).
$$

(11)

The entropy of a probability distribution measures the degree of randomness in the random variable. The entropy is used as a measure of uncertainty of the Dirichlet distributions of the map:

$$
U_t(x, y) = H(\text{Dir}(\beta_{t}^{x,y})), \forall x, y
$$

(12)

The last uncertainty measure is defined as the difference between the two classes of largest expected value. Let $p_{t,i}^{x,y}$ and $p_{t,j}^{x,y}$ be, respectively, the largest and second largest expected class posterior probabilities at map coordinates $(x, y)$. Then, we minimize:

$$
U_t(x, y) = \|p_{t,i}^{x,y} - p_{t,j}^{x,y}||.
$$

(13)

The predictions for the map at time $t+1$ being given by (8), the expected values of the uncertainty for all possible next gaze locations $(x', y')$ are computed through:

$$
U_{t+1}^{x',y'}(x, y) = \mathbb{E}[U_t(x, y) | B_{t+1}^{x',y'}(x, y)]
$$

(14)

Finally, we minimize over all possible next gaze directions $(x', y')$ the expected uncertainty accumulated over all map coordinates $(x, y)$:

$$
(x^*, y^*) = \arg\min_{x', y'} \sum_x \sum_y U_{t+1}^{x',y'}(x, y)
$$

(15)

III. RESULTS AND DISCUSSION

A. Experimental Setup

A set of 50 images was randomly chosen from the COCO 2017 validation dataset (80 object classes) as scenes in which to perform visual exploration. The employed Deep Object Detector [10] is pre-trained on the COCO 2017 training dataset. To validate our approach, we have incrementally implemented the pipeline illustrated in Fig. 4. The incremental implementation corresponds to the successive evaluation of each of the three key methods of the approach:

1) Foveal Observation Model: the parameters of each Dirichlet distribution are learned by employing an iterative method [11]; to analyse the effect of class classification scores calibration by the Foveal Observation Model in a simple classification task, the 50 images were foveated at randomly chosen focal points; the match between the detected bounding boxes and the ground-truth bounding boxes was considered valid for an Intersection over Union (IoU) with the ground-truth
information greater than 30%; then, for all matches, the classifications induced by the classifier scores (the class corresponding to the maximum classification score) were compared with the ground-truth, with and without the calibration performed by the Foveal Observation model; the entropy of the classification scores was also compared, in both calibration cases.

2) Data Fusion techniques: for each map cell \((x,y)\), there might be none to several detections at a given time step \(t\); Let \(I_{x,y}^{t-9}\) be the set of detections at instant \(t\) whose bounding boxes intersect the map cell \((x,y)\); at each time step \(t\), the map cell \((x,y)\) update is repeated for every detection belonging to the set \(I_{x,y}^{t-9}\); the calibrated classification scores by the Foveal Observation Model (1) serve as input to the four fusion techniques: Naïve Bayes update (2), Kaplan Update, Modified Kaplan’s rule (6) and sum rule (5); the next gaze direction was determined randomly instead of through Active Perception methods, in order to evaluate the raw performance of the fusion methods without considering gaze selection; the mean of the accuracy metrics across images was computed.

3) Active Perception techniques: the performance of the different acquisition functions was compared; each of the 50 images has been foveated 10 times, with the foveation points having been chosen by the various active perception techniques; the initial focal point was randomly chosen for each image, being the same for every technique; the detections at each iteration were used to update the information of the map using the Modified Kaplan update (6).

B. Experimental Results and Discussion

1) Validation of the Foveal Observation Model: In Fig. 5(a), the comparison of the classification performance with and without the Foveal Observation Model is illustrated, with the metrics values resulting from averaging over all the images. The accuracy lines in Fig. 5(a) are very similar, meaning that modelling the detections with the observation model does not generally lead to a drop in detection performance, proving the validity of the Foveal Observation Model. On the other hand, detection uncertainty is vastly improved, as shown in Fig. 5(b): low confidence scores output by the detector (i.e., detections significantly affected by the blur on the peripheries) typically have a high degree of entropy; when these scores are calibrated by the Foveal Observation Model, the entropy of the class confidence scores is much lower.

This is because the Foveal Observation Model tries to find the object class that better fits the distribution of the scores, accounting for the distance to the focal point (blur imposed by the foveation process), even if there is a big confusion among some of the classes, to present a more certain classification. Although this does not constitute an improvement in a one-step classification approach, as the lines in Fig. 5(a) are quite similar, it will still be useful for the multi-step classification approach that we are taking.

2) Validation of the Fusion techniques: In Fig. 6, one can see how the average accuracy evolves as new bounding boxes are detected (each bounding box corresponds to one classification). It is possible to note that every algorithm achieves a similar performance on the accuracy, except for the Naïve Bayes, where the performance is lower. Also, due to the drastic entropy reduction imposed by the Foveal Observation Model, the Naïve Bayes approach was greatly affected by most of the classes having a score closer to zero.

3) Validation of the Active Perception techniques: In Fig. 7, one can observe that the KL Divergence is the acquisition function achieving the highest F1-score. In Fig. 8, active gaze selection is compared with random gaze selection. The experiment is the same as before, but now the Modified Kaplan fusion method with KL divergence acquisition function is tested against all the other fusion methods which choose the next focal point randomly. One can immediately notice that the Active Perception method achieves higher F1-Score at almost every iteration than the other fusion algorithms with random gaze selection.

Another important aspect is the growth rate of the performance on classifying the objects on the image. Since the goal is to find and classify every object on the image in the least number of gaze shifts, analysing how fast the algorithm can detect and correctly classify most of the objects is a key factor. As one can see, choosing the next focal point by maximizing the predicted gain on the average KL divergence of the map, achieves an F1-Score around the third iteration that can not be surpassed by any of the methods that use random search.

Besides the improved growth rate, we can also see in Fig. 8 that choosing the next focal point by maximizing the KL Divergence Gain contributes to an overall performance improvement (on average) of around 2-3% on the F1-Scores after the 10 iterations of the experiment.

IV. Conclusions

In this work, we propose a methodology integrating a simulated foveal sensor, an object detector, an observation model, and data fusion and active perception techniques to perform visual exploration of a scene, in order to locate and correctly classify the objects in an image with the least number of gaze shifts. The main contribution of the paper is the Foveal Observation Model, which calibrates the scores of the Deep Object Detector to the foveal image topology in a fast training process without the need to retrain the deep object detector itself. We have shown that this calibration process significantly reduces the entropy of the classification score vectors. We embed the Foveal Observation model in several data fusion processes an evaluate performance both in random search and with active search. The proposed method is able to improve the F1-scores by 2-3 pp after a few gaze shifts, while the active perception method also achieves a score higher than the best achieved when randomly searching.
Fig. 5. Classification entropy employing (blue) and without employing (red) the Foveal Observation Model.

Fig. 6. Accuracy versus number classifications at each map location, for the fusion techniques considered.

Fig. 7. Comparison between the F1-Scores using the three different acquisition functions.

Fig. 8. F1-Score comparison of the Modified Kaplan using the acquisition function "KL Divergence Gain" (red), against the ones choosing the focal point randomly.

ACKNOWLEDGMENTS
This work was supported by FCT through projects LARSyS-UIDB/50009/2020, HAVATAR-PTDC/EEI-ROB/1155/2020, and SHIFT-HRI-CMU/TIC/0026/2021.

REFERENCES
[1] J. H. Krantz, “The Stimulus and Anatomy of the Visual System,” in Experiencing Sensation and Perception, ch. 3, pp. 3.1 – 3.36, 2012.
[2] 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.
[3] A. Aydemir, K. Sjöo, J. Folkesson, A. Pronobis, and P. Jensfelt, “Search in the real world: Active visual object search based on spatial relations,” 2011 IEEE International Conference on Robotics and Automation, pp. 2818–2824, 2011.
[4] A. Aydemir, A. Pronobis, M. Gobelbecker, and P. Jensfelt, “Active visual object search in unknown environments using uncertain semantics,” IEEE Transactions on Robotics, vol. 29, no. 4, pp. 986–1002, 2013.
[5] A. F. Almeida, R. Figueiredo, A. Bernardino, and J. Santos-Victor, “Deep Networks for Human Visual Attention: A Hybrid Model Using Foveal Vision,” in ROBOT 2017: Third Iberian Robotics Conference (A. Ollero, A. Sanfeliu, L. Montano, N. Lau, and C. Cardeira, eds.), (Cham), pp. 117–128, Springer International Publishing, 2018.
[6] 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), pp. 101–106, 2018.
[7] M. Grotz, T. Habra, R. Ronse, and T. Asfouri, “Autonomous view selection and gaze stabilization for humanoid robots,” in 2017 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS), pp. 1427–1434, 2017.
[8] 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, no. 2, pp. 459–476, 2018.
[9] L. M. Kaplan, S. Chakraborty, and C. Bisiklia, “Fusion of classifiers: A subjective logic perspective,” in 2012 IEEE Aerospace Conference, pp. 1–13, 2012.
[10] J. Redmon and A. Farhadi, “YOLOv3: An Incremental Improvement,” 2018.
[11] T. Minka, “Estimating a Dirichlet distribution,” Technical report, M.I.T., 2000.