Robust top-down and bottom-up visual saliency for mobile robots using bio-inspired design principles

Uziel Jaramillo-Avila[1], Jonathan M. Aitken[2], Kevin Gurney[3] and Sean R. Anderson[4]

Abstract— **Modern camera systems in robotics tend to produce overwhelming amounts of visual information due to their high resolutions and high frame rates. This raises a fundamental question of how robots should focus attention on a region of the visual scene, and how they should process information in the periphery. This is particularly an issue for mobile robots, where the computational resources of low-power embedded computing boards tend to be much less than for workstations. In this paper, we look to biological design in the primate brain for inspiration on how to solve this problem. We develop a novel computational fusion of bottom-up and top-down visual saliency information. The bottom-up saliency is produced using standard colour, intensity, and motion image processing methods. The top-down saliency is produced using a deep convolutional neural network for object detection and recognition, with foveated images for computational efficiency. Regions of attention are obtained using a computational model of the basal ganglia, thought to be involved in optimal decision making, which improves robustness to noise. The model of the basal ganglia is based on the multi-hypothesis sequential probability ratio test (MSPRT). The visual saliency scheme is evaluated on omnidirectional video feed highlighting a proximity to human behaviour.**

I. Introduction

There is an overwhelming amount of visual information typically acquired by vision sensors in robotics, due to the significant increases in camera resolution and, in recent years, the use of omnidirectional cameras. This raises a fundamental question of how robots should focus their attention on a region of the visual scene, and how they represent and process information in the periphery. In biology, visual saliency is used to shift attention, where the fovea, a dense region of photoreceptors, is redirected to the current point of interest [1]. The periphery is sampled using a much less dense region of photoreceptors, where density falls off exponentially away from the fovea [2]. This is highly computationally efficient: it enables the brain to compactly analyse the foveated region of attention in high detail, whilst also monitoring the periphery with much lower computational load. The aim of this paper is to model the core, functional components of the human visual saliency system using machine learning analogs, in order to improve the computational efficiency of robot vision systems, and to expand on the problem of the computational burden of vision in robotics: in recent years deep convolutional neural networks (DCNNs) have become dominant for tasks such as object detection and recognition. These algorithms work impressively well, surpassing conventional methods on competition problems in image recognition [3]. However, they incur a major computational cost, and are often implemented on workstations with high-end Graphics Processing Units (GPUs). These DCNN systems are often not particularly well suited for battery powered mobile robots.

In previous work we have demonstrated that applying a foveated image transform can speed up DCNN object detection and recognition in low power embedded GPUs by up to a factor of $4\times$, with only a marginal drop in detection and recognition performance in the foveated region [4]. The foveated image transform in effect, zooms the area of attention with dense sampling and reduces sampling in the periphery. This type of foveated transform has been used elsewhere in computer vision to similarly improve computational efficiency [5]. The use of foveated vision raises the challenge of directing the fovea towards a region of interest. This is why it is necessary to use a visual saliency system in conjunction with a foveated vision system.

We present a bioinspired visual saliency system that is modelled at a high, functional level, on the human system. This scheme is based on a standard model of human visual saliency where bottom-up saliency is additively combined with weighted top-down saliency to redirect the fovea [6]. We extend here primarily by introducing a computational model of the basal ganglia. In the human brain it is thought that the basal ganglia act as a central device that accumulates saliency evidence over time to take robust decisions [7]. We use a simple computational model of the basal ganglia [8], [9] that has been shown to be related to the multi-hypothesis sequential probability ratio test (MSPRT) for decision making [10]. This MSPRT is a computationally lightweight model of basal ganglia decision making function, not a detailed biophysical model such as in [11], but well suited to robotic systems for its computational efficiency. The decision making algorithm uses evidence accumulation of the saliency over time and threshold testing to decide which region is most salient.

II. Methods

The complete model of visual attention has the following components. Each frame from the camera is processed by

[1] Uziel Jaramillo-Avila is with the Department of Computer Science of the Center for Mathematical Research, Zacateacas, Mexico. uziel.jaramillo@cimat.mx

[2] Jonathan M. Aitken is with the Department of Automatic Control and Systems Engineering, University of Sheffield. Sheffield, United Kingdom. jonathan.aitken@sheffield.ac.uk

[3] Kevin Gurney is with the Department of Psychology, University of Sheffield. Sheffield, United Kingdom. k.gurney@sheffield.ac.uk

[4] Sean R. Anderson is with the Department of Automatic Control and Systems Engineering, University of Sheffield. Sheffield, United Kingdom. s.anderson@sheffield.ac.uk

a bottom-up pathway to extract low-level features such as orientations, colour and intensity, as well as movement, emulating low-level processing in the thalamus and visual cortex. The image undergoes a foveated transform that is processed in a top-down pathway by a DCNN object detection and recognition system, emulating the 'where' and 'what' high-level processing in the ventral and dorsal pathways of the brain. Thus, task relevant information is embedded in terms of the discrete label categories that the DCNN is trained to detect. The resulting saliency maps of the top-down and bottom-up pathways are additively fused using a weighted average. The resulting saliency of each image region is transmitted to the MSPRT algorithm to select the most salient region, emulating basal ganglia function. The selected region modifies a final sensorimotor map, emulating the sensorimotor map in superior colliculus that directs gaze. The peak in the sensorimotor map defines the direction of the fovea. A fast pathway interrupt is also included that links low-level image processing movement detection to the final sensorimotor map.

The system is tested and evaluated on an omnidirectional dataset which includes human eye fixations (location and duration) [12]. This type of videos are typically preserved in an equirectangular projection covering an horizontal field of view of 360° and of 180° vertically. It would be computationally infeasible to process this entire projection in a DCNN, and selecting a sub-region from the field of view to transform to rectilinear coordinates is far from a trivial question, and represents a large computational bottleneck for robotic systems. We use our visual saliency scheme to efficiently tackle this problem: where regions of the omnidirectional image are only processed by the DCNN once enough saliency evidence is accumulated by the MSPRT algorithm for a given region.

A. *Visual saliency scheme using bioinspired design principles*

Given the large number of models that are present in the literature aiming to establish a functional scheme of the brain and visual processing, it is difficult to present an unanimous agreement of its connectivity. One prominent theory, often referred to as the "two-streams hypothesis" [13], is that the brain has two distinct processing pathways for visual and auditory information: the dorsal and ventral streams. The former focuses on where objects are located in space and in action planning [14], often called "where" stream. The latter stream is associated with object categorization, also known as the "what" stream. Fig. 1 presents a diagram of the analogy between our current model and the connectivity of regions in the the macaque brain and their functionality.

1) Bottom-up saliency using low-level image processing: Different approaches have been proposed to adapt saliency implementations to omnidirectional cameras. Most of these are inherently computationally heavy, e.g. by (i) using deep convolutional neural networks, (ii) making several rectilinear projections from the image, obtaining their saliency and fusing them back to an equirectangular one, or (iii) calculating several complementary approaches and then aggregating them into a global saliency. Here we opt to use for omnidirectional frames the same bottom-up saliency method based on the computational model presented in [16], i.e. Vocus2, by making the adaptation from having a prior, or bias (λ), towards the equator instead of the central point, along which the vast majority of human fixations fall [17]. Using Vocus2 [16], the bias is defined at $\lambda_{eq} = 5 \times 10^{-5}$. The rest of the major parameters are set to: $\sigma_{center} = 1$, $\sigma_{surround} = 2$, four stop layers and the use of the arithmetic mean for feature and conspicuity fusion, which leads to lightweight bottom-up saliency for the omnidirectional camera. To implement bottom-up saliency, we used the Vocus2 C++ toolbox [18], where we ported it as a Python library to integrate with the rest of the visual saliency system.

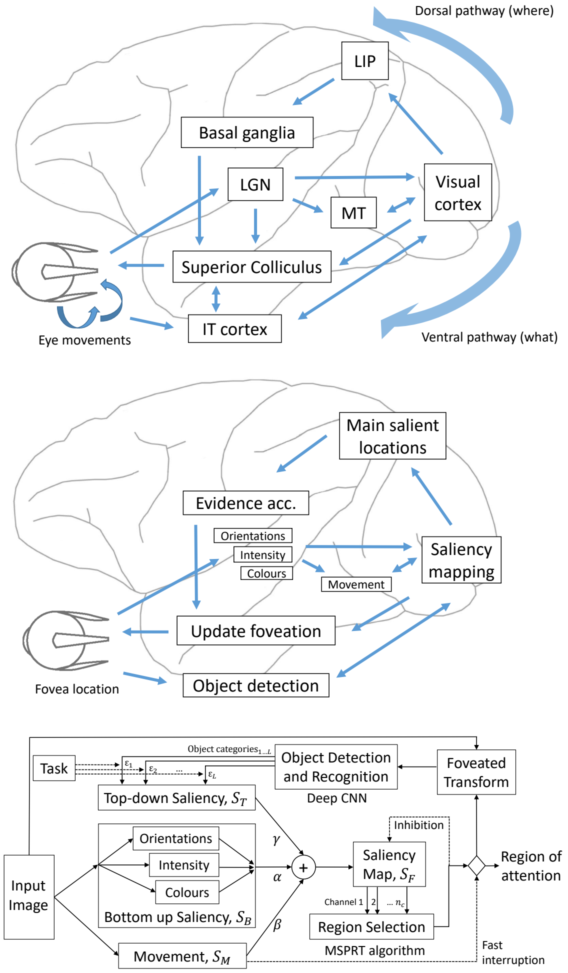

Fig. 1. Top: Brain regions involved in visual processing emphasising the dorsal and ventral pathways: the where and what. It has been proposed that a bottom-up saliency map is formed in the primary visual cortex [15]. The basal ganglia is thought to accumulate evidence from different parts of the cortex, including those associated with eye movements, such as the lateral intraparietal area (LIP) and the frontal eye field (FEF). Middle: Functional diagram of visual processing. Bottom: Visual saliency block diagram.

Bottom-up saliency fits into our approach by first making a global bottom-up estimation using a normally downscaled frame. This greatly speeds up the process (we use a down-

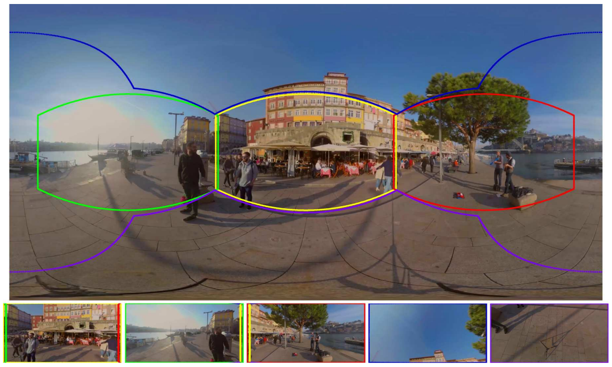

Fig. 2. **(Top)** Example of an equirectangular frame, at 3840×1920 pixel resolution. **(Bottom, from left to right)**, **(a)** Rectilinear projection centered in [λ (horizontal coordinate system), ϕ (vertical coordinate system)] = [0.5, 0.5], with a resolution of 768x384 pixels, representing about $114°$ in latitude, as illustrated by the yellow borders in the equirectangular frame **(b)** Projection centered at [0.2, 0.5], green region in the equirectangular frame **(c)** Region centered in [0.8, 0.5] illustrated by the red border **(d)** Centered in [0.5, 0.1] and bordered in blue **(e)** Centered in [0.5, 0.9] and bordered in purple.

scale ratio of 5:1, to bring the 4K resolution frame to Wide-VGA). Five regions are used as channels for evidence accumulation; the central one (where the current fixation is positioned), to its left, right, up and down, as exemplified in Fig. 2.

2) Foveated image transform: Photoreceptor density and cortical magnification factor have been well studied in the biological domain [19], [20], which can be used to inform computational models. A number of different computational methods have been developed to transform a uniformly sampled digital image into a foveated image, including the log-polar transform, the reciprocal wedge-transform and Cartesian foveated geometry. The advantage of the Cartesian log-spaced sampling is that it distorts the original image less than, e.g. a log-polar transform, and therefore has the key benefit of enabling the use of transfer learning to speed-up the training of the DCNN (i.e. initialising the DCNN weights using a network pre-trained on uniformly sampled images).

The method used here [4] re-samples the uniform digital image of size $N_x \times N_y$ pixels, to a new size of $n_x \times n_y$ pixels with log-spacing, so that for the upper right quadrant of the image with the fovea centred on (x_0, y_0) we have sample locations,

$$x_k = \exp(k\Delta_x) \quad \text{for } k = 0, \ldots, n_x/2 \tag{1}$$

$$y_k = \exp(k\Delta_y) \quad \text{for } k = 0, \ldots, n_y/2 \tag{2}$$

where

$$\Delta_x = 2n_x^{-1} \log(N_x/2) \tag{3}$$

$$\Delta_y = 2n_y^{-1} \log(N_y/2) \tag{4}$$

An illustration of the foveated transform on omnidirectional images in shown in Fig. 3, where first, a sub-region of the full image is selected, which is transformed using the equirectilinear projection following the method in [21]. Subsequently, the top salient location of that region is taken as the central point to foveate using the transformation described above.

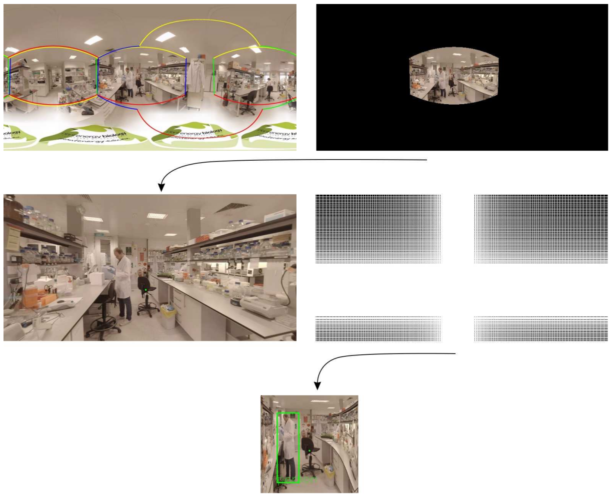

Fig. 3. Top left: equirectangular sample frame (at 3840×1920 resolution). Top right: masked region of interest, selected using the MSPRT algorithm. Middle left: rectilinear projection of the area of interest (at 768×384 resolution) - a small dot in the chair near the center indicates it to be the most salient point of that region. Middle right: selected rows and columns to foveate the image. Bottom: foveated frame at 256×256 resolution, with the bounding box of the TinyYolo [22] prediction, run at the same resolution.

3) Target selection using evidence: The overall, fused saliency map S_F is calculated as the weighted average of the movement saliency map S_M, bottom-up map S_B and top-down map S_T,

$$S_F = \alpha S_B + \beta S_M + \gamma S_T \tag{5}$$

where α, β and γ are the weights, and $\alpha + \beta + \gamma = 1$. The weights can be tuned to adjust the influence of each saliency map. Here we set $\alpha = \beta = 0.4$ and $\gamma = 0.2$. The top-down weight γ is tuned to be smaller than the other weights because the top-down influence consists of filled rectangular bounding boxes for every object detected with confidence ≥ 0.5, thus the bounding box tends to be larger than the actual detected object, necessitating a reduction in the weight, otherwise the object becomes over-weighted.

We use the MSPRT decision making model of the basal ganglia [8], to perform region selection in the fused saliency map S_F. The MSPRT algorithm takes discrete channels of salience as input and accumulates evidence for each channel until a threshold is reached and a decision is made. To implement the MSPRT algorithm we require a small number of discrete input channels, but the fused saliency map, S_F, consists of many contiguous pixels of the same size as the input image. Therefore, we use a small number, n_c, of candidate regions in S_F as the discrete input channels. Here we use fixed regions corresponding to locations from an equirectangular projection, and a binary mask, M, to extract the region of interest, which is illustrated in Fig. 4.

To implement the MSPRT model of basal ganglia for region selection, the fused saliency map, S_F, is divided into n_c discrete regions corresponding to the number of

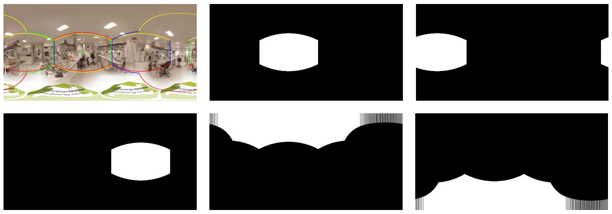

Fig. 4. Sample of region masks M_i; (from left to right) **(a)** Sample frame, **(b)** mask M_0 for region centered at $[\lambda, \phi]$, **(c)** mask M_1 for region $[\lambda - 0.3, \phi]$, **(d)** mask M_2 for $[\lambda + 0.3, \phi]$, **(e)** mask M_3 for region $[\lambda, \phi - 0.4]$, **(f)** mask M_4 for region $[\lambda, \phi + 0.4]$. The borders of the M_i masks are pre-saved on a lookup table (LUT) for all possible locations with a 0.01 resolution for λ and ϕ.

input channels, and a channel i is selected (disinhibited in the context of the basal ganglia) if the output O_i, from accumulated evidence in channel i, crosses a fixed threshold Θ, where

$$O_i(T) = -y_i(T) + \log \sum_{j=1}^{n_c} \exp\left(y_j(T)\right) \text{ for } i = 1, \ldots, n_c \tag{6}$$

where $y_i(T) = gY_i(T)$ (g is a scaling parameter set to $g = 1$ here), and where

$$Y_i(T) = \sum_{t=t_0}^{T} s_i(t) \text{ for } i = 1, \ldots, n_c \tag{7}$$

where $Y_i(T)$ is the accumulated evidence in channel i, between frames $[t_0, T]$ and s_i is a scalar value of saliency obtained from summing over a region $S'_{F,i}$, which is the ith region of S_F extracted using a binary mask M_i. We use $t_0 = T - 25$, so that $Y_i(T)$ takes as evidence the stimuli present in the last ~ 1 second (assuming a framerate of 25-30 fps). Note that the threshold Θ is a hyperparameter that must be tuned to give effective performance, typically in a speed-accuracy sense (i.e. faster decisions with less accuracy or slower decisions with more accuracy).

In order to incorporate an inhibition of return (IOR) influence, which partially impairs a winning region once it crosses the threshold Θ, we add an influencing factor Ω_i to each evidence channel in Eqn. (7),

$$Y_i(T) = \Omega_i \sum_{t=t_0}^{T} s_i(t) \tag{8}$$

where $\Omega_i = 1$ for a channel that has not been selected and $\Omega_i = 0.5$ for a channel that has just been selected, returning linearly over time to $\Omega_i = 1$ (over an interval of 10 time-steps here).

B. *Omnidirectional image processing*

Transforming from an equirectangular to rectilinear camera projection (where straight lines are displayed undistorted), and then performing object detection, is a computationally heavy operation. But even if the former can be considered simply an implementation requirement, the latter (or both of them together) is arguably the biggest bottleneck, in a comparable sense in which the human eye can only target the fovea an thus identify objects with certainty in a very small region. It is also equivalent in how the eye and head movement can not be guided by the immediate prominent stimulus in every instant, given movement constraints, both for a human and for a mobile robot.

Functionally, projecting from an equirectangular representation to rectilinear is restricted to a section of about 0.3 of the equirectangular frame along the equator. Since 1.0 of the figure is the full 360°, a region of 114° represents; $114 \div 360 = 0.3166$. With help of the [21] Toolbox, a sub-region of this size of the 4K equirectangular frame is transformed to a frame in the range of Wide-VGA resolution (768×384 pixels). Although for certain applications, this image size might be enough to run through object detection, depending on the GPU capability, we sustain the hypothesis that taking it a step further and foveating in the most promising region of interest is advantageous, since it enables the system to run on low power GPUs, which is here exemplified using the TinyYolo [22] DCNN at 224×224 resolutions.

The information flow described in the previous paragraph also permits to sustain the analogy to the human vision system, which is continually faced with the same dilemma. The average human field of view has an approx. 210° forward-facing horizontal range, and a 150° vertical one [23]. Although the high acuity region, known as fovea centralis, is concentrated in a region of about 5°, and the foveola in 1°, with the highest visual acuity. Hence, humans constantly chose from an unbounded, often unstructured environment, the region to fit into our field of view (by body/head movements) and subsequently, fixate our fovea into small targets (by eye movements), leaving the majority of the region in our peripheral vision. Three consecutive 114° regions along the latitude would cover most of the 360° field of view, depending on where the central point is located. Equivalently three 57° (conserving the 2:1 image ratio) regions cover most of the 180° longitude field of view.

As exemplified in Fig. 2, with $[\lambda_0, \phi_0] = [0.5, 0.5]$ (i.e. the center of the frame) the area that can be easily transformed to a rectilinear representation is delimited by the yellow border. Moving the focus point to $[\lambda_0 - 0.3, \phi_0]$ and $[\lambda_0 + 0.3, \phi_0]$ would produce the regions delimited by green and red borders, respectively. A similar process can be done to transform the blue and purple bordered regions by centering in $[\lambda_0, \phi_0 - 0.4]$ and $[\lambda_0, \phi_0 + 0.4]$ accordingly. This is a coarse rule where, due to the equirectangular distortion, a different amount of overlap is caused depending on where the central point is. Even with this overlap, most of the scene is covered by the five regions, particularly along the equator, where the vast majority of fixation will fall [12]. By bordering the omnidirectional scene in such a way, it can be approached with a divide-and-conquer outlook. By first exploring the saliency of the full scene, and transforming to rectilinear only a subsection once enough saliency has been accumulated for it.

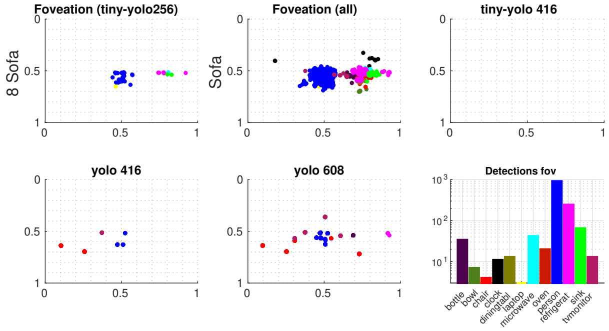

Fig. 5. (From left to right) Example of object detections performed at different network resolutions and for the foveated regions, for one of the videos on the omnidirectional dataset; **(a)** Detections on foveated frames, mapped back to their equirectangular location, using the TinyYolo network at 256 × 256 resolution, **(b)** Repeating the detection on foveated frames at the same resolution, following the observation position of all participants in the omnidirectional dataset **(c)** No detections were possible when tested on every equirectangular frame using TinyYolo at 416 × 416 resolution **(d)** Detections using yoloV3 at 416 × 416 on omnidirectional frames **(e)** Detections using yoloV3 at 608 × 608 on omnidirectional frames **(f)** Graph of the main detected objects on the foveated frames.

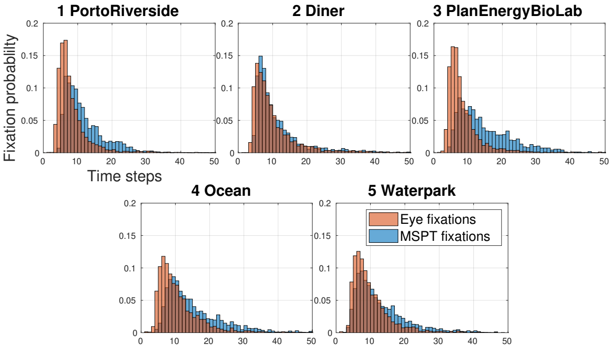

Fig. 6. Histogram of duration of the fixations in the first five videos of the [12] dataset. For each video, the horizontal axis represents the number of frames that the fixation lasts, and the vertical the probability that the fixation falls within that range.

C. Experimental Data: Eye movements for 360° videos

The Omnidirectional dataset [12] that was used consists of 19 videos, each of 20 seconds in equirectangular format. With frame rates between 24 - 30 fps, observed by an average of 49.63 participants (median = 50), it amounts to over half a million fixated-on frames. With this data, we aim to get an overall behaviour comparison between human fixations and our proposed visual saliency system, whilst also demonstrating how the regions of interest selected by the system (and foveated on) proffer a good strategy to select visual information to run through a light-weight DCNN object detection system, without it requiring any special modification or retraining.

As a point of interest, the distribution of the duration of fixations (in terms of frames) is compared between the dataset of human fixations and those produced by our MSPRT system. To ensure that both systems are looking at similar data (since neither a human observer nor our system look at the whole 360° at a time), the fixation point $[\lambda, \phi]$ from the observers in the dataset is followed in every frame, also deriving the rest of the I channels from it. The MSPRT threshold cross $O_i(n) < \Theta$, triggers a new foveal transformation and CNN detection feedback.

III. Results

As a first step, we show the object detection performance empirically on the omnidirectional dataset, showing a big advantage to foveating over different segments of the frame, in contrast to trying to perform detection on the full frame. Fig. 5 illustrates how for YoloV3 and TinyYolo at lower resolutions, the system fails to make detections, partially because of the omnidirectional distortion (although most objects are near the equator, where the distortion is less) and partially because the wide angle representation ends up making most objects too small to be found by a lightweight object detector. One important point to note is that by varying to location of the fovea, we obtain different but complementary detections, in contrast to performing detection at every frame at a high resolution, where the drawbacks will be persistent.

The Kolmogorov-Smirnov Chi-Square *K-S* and Chi-Square χ^2 are two *goodness-of-fit* test values (for statistical hypothesis testing), that are commonly used to analyse visual saliency algorithms. Given the histograms of duration of fixations, for the ground truth and our implementation, H_1 represents the distribution of duration of human fixations, and H_2 represents the duration of MSPRT fixation, both with a bin-width equal to 1, and considering $m = 45$ bins. Then the *K-S* and χ^2 are defined as [24],

$$\text{K-S} = \max_{1 \leq i \leq m} \left| \sum_{i=1}^{m} H_1(i) - \sum_{i=1}^{m} H_2(i) \right| \tag{9}$$

$$\chi^2 = \sum_{i=1}^{m} \frac{(H_1(i) - H_2(i))^2}{H_1(i) + H_2(i)} \tag{10}$$

To illustrate how the system compares to human fixation behaviour, a simple first test is to study the histograms of the duration of fixations, using the human recorded data provided for the omnidirectional dataset [12]. Fig. 6 shows an histogram of their duration for the first five videos of the dataset. Tables I compile the Kolmogorov-Smirnov (*K-S*) and Chi-Square (χ^2) for the first 15 videos. With total average values of $\overline{\text{K-S}} = 0.269$ and $\overline{\chi^2} = 0.284$, illustrating the similarity for most videos between both histograms, noting that if $H_1 = H_2$, then K-S $= \chi^2 = 0$. Evidently the proximity between the histograms can be brought closer for each video, by parameter tuning, if it were the sole goal. Another relevant aspect to evaluate is how often top salient locations are fixated-on, both in contrast to human fixations and to a simple Winner-Takes-All (WTA) rule. In Fig. 7 we illustrate how the WTA fixates on a small number of prominent locations (the larger the circle is drawn represents the number of times that the location is fixated-on), while the human data and the MSPRT and more broadly distributed.

Video	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
K-S	0.149	0.306	0.254	0.280	0.128	0.463	0.457	0.252	0.224	0.141	0.219	0.324	0.174	0.255	0.059
χ^2	0.113	0.342	0.298	0.275	0.086	0.563	0.556	0.243	0.159	0.185	0.241	0.333	0.159	0.181	0.045

TABLE I

KOLMOGOROV-SMIRNOV (*K-S*) AND CHI-SQUARE (χ^2) GOODNESS-OF-FIT TESTS FOR STIMULUS VIDEOS 1 TO 14.

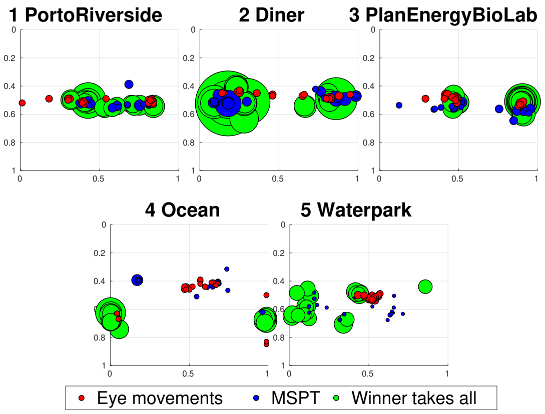

Fig. 7. Representation of the top 20 fixations for the first five stimulus videos. The size of the circle represents how many times each location appears, with a $[\lambda, \phi]$ precision of $[0.01, 0.01]$. The eye fixations from the dataset (illustrated in red) are not too often dominated by a single location, behaviour which is much more closely imitated by an evidence accumulation system, like the presented MSPRT, than by a WTA rule in every frame.

IV. CONCLUSIONS

In this paper, we have developed a novel robust visual saliency system for mobile robots, to select a region of interest in the visual scene, which fuses top-down and bottom-up saliency, and performs region selection using a bioinspired evidence accumulation algorithm related to basal ganglia decision making. The key components of the system are: top-down saliency using a foveated image transform with DCNNs for fast object detection and recognition (the where and what), which is combined with biasing by task relevant information; bottom-up saliency using standard low-level image processing from intensity, colour and orientation maps; movement saliency combined with a fast path interrupt to bypass the evidence accumulation algorithm and rapidly direct attention towards potential hazards. The results demonstrate that the visual saliency system works effectively to select regions of attention; that is speeds up DCNN processing; that the system emulates human visual saliency more closely than schemes that use a winner-take-all decision making rule; and that the system is more robust to noise.

ACKNOWLEDGMENT

Most of this work was done while Uziel Jaramillo-Avila was at The University of Sheffield, funded by CONACyT.

REFERENCES

[1] D. H. Hubel, J. Wensveen, and B. Wick, *Eye, brain, and vision*. Scientific American Library New York, 1995.

[2] D. Purves, R. Cabeza, S. A. Huettel, K. S. LaBar, M. L. Platt, M. G. Woldorff, and E. M. Brannon, *Neuroscience*. Sunderland, 2004.

[3] O. Russakovsky, J. Deng, H. Su, J. Krause, S. Satheesh, S. Ma, Z. Huang, A. Karpathy, A. Khosla, M. Bernstein, A. C. Berg, and L. Fei-Fei, "Imagenet large scale visual recognition challenge," *International Journal of Computer Vision*, vol. 115, no. 3, pp. 211–252, 2015.

[4] U. Jaramillo-Avila and S. R. Anderson, "Foveated image processing for faster object detection and recognition in embedded systems using deep convolutional neural networks," in *Lecture Notes in Computer Science*. Springer, 2019.

[5] E. Akbas and M. P. Eckstein, "Object detection through search with a foveated visual system," *PLoS Computational Biology*, vol. 13, no. 10, p. e1005743, 2017.

[6] U. Jaramillo-Avila, J. M. Aitken, and S. R. Anderson, "Visual saliency with foveated images for fast object detection and recognition in mobile robots using low-power embedded GPUs," in *2019 19th International Conference on Advanced Robotics (ICAR)*. IEEE, 2019, pp. 773–778.

[7] P. Redgrave, T. J. Prescott, and K. Gurney, "The basal ganglia: a vertebrate solution to the selection problem?" *Neuroscience*, vol. 89, no. 4, pp. 1009–1023, 1999.

[8] R. Bogacz and K. Gurney, "The basal ganglia and cortex implement optimal decision making between alternative actions," *Neural computation*, vol. 19, no. 2, pp. 442–477, 2007.

[9] N. F. Lepora and K. N. Gurney, "The basal ganglia optimize decision making over general perceptual hypotheses," *Neural Computation*, vol. 24, no. 11, pp. 2924–2945, 2012.

[10] C. W. Baum and V. V. Veeravalli, "A sequential procedure for multihypothesis testing," *IEEE Transactions on Information Theory*, vol. 40, no. 6, 1994.

[11] K. Gurney, T. J. Prescott, and P. Redgrave, "A computational model of action selection in the basal ganglia. I. A new functional anatomy," *Biological Cybernetics*, vol. 84, no. 6, pp. 401–410, 2001.

[12] E. J. David, J. Gutiérrez, A. Coutrot, M. P. Da Silva, and P. L. Callet, "A dataset of head and eye movements for 360 videos," in *Proceedings of the 9th ACM Multimedia Systems Conference*, 2018, pp. 432–437.

[13] M. A. Goodale and A. D. Milner, "Separate visual pathways for perception and action," *Trends in neurosciences*, vol. 15, no. 1, 1992.

[14] N. Kruger, P. Janssen, S. Kalkan, M. Lappe, A. Leonardis, J. Piater, A. J. Rodriguez-Sanchez, and L. Wiskott, "Deep hierarchies in the primate visual cortex: What can we learn for computer vision?" *IEEE transactions on pattern analysis and machine intelligence*, vol. 35, no. 8, pp. 1847–1871, 2012.

[15] Z. Li, "A saliency map in primary visual cortex," *Trends in cognitive sciences*, vol. 6, no. 1, pp. 9–16, 2002.

[16] S. Frintrop, T. Werner, and G. Martin Garcia, "Traditional saliency reloaded: A good old model in new shape," in *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition*, 2015.

[17] Y. Rai, J. Gutiérrez, and P. Le Callet, "A dataset of head and eye movements for 360 degree images," in *Proceedings of the 8th ACM on Multimedia Systems Conference*, 2017, pp. 205–210.

[18] S. Frintrop, "Saliency System VOCUS2," [online] http://pages.iai.uni-bonn.de/frintrop_simone/vocus2.html, 2015, accessed: 03-05-2020.

[19] E. L. Schwartz, "Spatial mapping in the primate sensory projection: analytic structure and relevance to perception," *Biological cybernetics*, vol. 25, no. 4, pp. 181–194, 1977.

[20] S. W. Wilson, "On the retino-cortical mapping," *International Journal of Man-Machine Studies*, vol. 18, no. 4, pp. 361–389, 1983.

[21] N. Mutha, "Equirectangular-toolbox," https://github.com/NitishMutha/equirectangular-toolbox, 2017.

[22] J. Redmon and A. Farhadi, "Yolov3: An incremental improvement," *arXiv preprint arXiv:1804.02767*, 2018.

[23] H. M. Traquair, *An introduction to clinical perimetry*, 5th ed. London: Kimpton, 1946.

[24] K. Meshgi, "Histogram of color advancements," 2014.