Third Iberian Robotics Conference (ROBOT)
Seville, Spain, 22-24 November 2017

Deep Networks for Human Visual Attention: A hybrid model using foveal vision

Ana Filipa Almeida, Rui Figueiredo, Alexandre Bernardino and José Santos-Victor

Institute for Systems and Robotics, Instituto Superior Técnico, Universidade de Lisboa, Lisbon, Portugal
ana.j.almeida@ist.utl.pt,
{ruifigueiredo,alex,jasv}@isr.tecnico.ulisboa.pt

Abstract. Visual attention plays a central role in natural and artificial systems to control perceptual resources. The classic artificial visual attention systems uses salient features of the image obtained from the information given by predefined filters. Recently, deep neural networks have been developed for recognizing thousands of objects and autonomously generate visual characteristics optimized by training with large data sets. Besides being used for object recognition, these features have been very successful in other visual problems such as object segmentation, tracking and recently, visual attention. In this work we propose a biologically inspired object classification and localization framework that combines Deep Convolutional Neural Networks with foveal vision. First, a feed-forward pass is performed to obtain the predicted class labels. Next, we get the object location proposals by applying a segmentation mask on the saliency map calculated through a top-down backward pass. The main contribution of our work lies in the evaluation of the performances obtained with different non-uniform resolutions. We were able to establish a relationship between performance and the different levels of information preserved by each of the sensing configurations. The results demonstrate that we do not need to store and transmit all the information present on high-resolution images since, beyond a certain amount of preserved information, the performance in the classification and localization task saturates.

Keywords: Computer vision, deep neural networks, object classification and localization, space-variant vision, visual attention.

1 Introduction

The available human brain computational resources are limited, therefore it is not possible to process all the sensory information provided by the visual perceptual modality. Selective visual attention mechanisms are the fundamental mechanisms in biological systems, responsible for prioritizing the elements of the visual scene to be attended. Likewise, an important issue in many computer vision applications requiring real-time visual processing, resides in the involved

computational effort [1]. Therefore, in the past decades, many biologically inspired attention-based methods and approaches, were proposed with the goal of building efficient systems, capable of working in real-time. Hence, attention modeling is still a topic under active research, studying different ways to selectively process information in order to reduce the time and computational complexity of the existing methods.

Nowadays, modeling attention is still challenging due to the laborious and time-consuming task that is to create models by hand, trying to tune where (regions) and what (objects) the observer should look at. For this purpose, biologically inspired neural networks have been extensively used, since they can implicitly learn those mechanisms, circumventing the need of creating models by hand.

Our work is inspired by [2] which proposed to capture visual attention through feedback Deep Convolutional Neural Networks. Similarly in spirit, we propose a biologically inspired hybrid attention model, that is capable of efficiently recognize and locate objects in digital images, using human-like vision. Our method comprises two steps: first, we perform a bottom-up feed-forward pass to obtain the predicted class labels (detection). Second, a top-down backward pass is made to create a saliency map that is used to obtain object location proposals after applying a segmentation mask (localization). The main contributions of this paper are the following: first, we evaluate the performance of our methodology for various well-known Convolutional Neural Network architectures that are part of the state-of-the-art in tasks of detection and localization of objects when combined with multi-resolution, human-inspired, foveal vision. Then, we establish a relationship between performance and the different levels of information preserved by foveal sensing configurations.

The remainder of this paper is organized as follows: section 2 overviews the related work and some fundamental concepts behind the proposed attentional framework. In section 3.1, we describe in detail the proposed methodologies, more specifically, a theoretical explanation of an efficient artificial foveation system and a top-down, saliency-based mechanism for class-specific object localization. In section 4, we quantitatively evaluate the our contributions. Finally, in section 5, we wrap up with conclusions and ideas for future work.

2 Background

The proposed object classification and localization framework uses several biologically inspired attention mechanisms, which include space-variant vision and Artificial Neural Networks (ANN). As such, in the remainder of this section we describe the fundamental concepts from neuroscience and computer science on which the proposed framework is based.

2.1 Space-variant Vision

In this work, we propose to use a non-uniform distribution of receptive fields that mimics the human eye for simultaneous detection and localization tasks.

Unlike the conventional uniform distributions which are typical in artificial visual systems (e.g. in standard imaging sensors), the receptive field distribution in the human retina is composed by a region of high acuity – the fovea – and the periphery, where central and low-resolution peripheral vision occurs, respectively [3].

The central region of the retina of the human eye named fovea is a photoreceptor layer predominantly constituted by cones which provide localized high-resolution color vision. The concentration of these photoreceptor cells reduce drastically towards the periphery causing a loss of definition. This space-variant resolution decay is a natural mechanism to decrease the amount of information that is transmitted to the brain (see Figure 2). Many artificial foveation methods have been proposed in the literature that attempt to mimic similar behavior: geometric method [4], filtering-based method [5] and multi-resolution methods [6].

2.2 Deep Convolutional Neural Networks

Deep neural networks are a subclass of Artificial Neural Networks (ANN) and are characterized by having several hidden layers between the input and output layers. The deep breakthrough occurred in 2006 when researchers brought together by the Canadian Institute for Advanced Research (CIFAR) were capable of training networks with much more layers for the handwriting recognition task [7].

As far as visual attention is concerned, the most commonly used are the Convolutional Neural Networks (CNN), that are feed-forward Deep ANN that take into account the spatial structure of the inputs. These, have the ability to learn discriminative features from raw data input and have been used in several visual tasks like object recognition and classification.

A CNN is constituted by multiple stacked layers that filter (convolve) the input stimuli to extract useful and meaningful information depending on the task at hand. These layers have parameters that are learned in a way that allows filters to automatically adjust to extract useful information without feature selection so there is no need to manually select relevant features. In this work we study the performance of state-of-the-art CNN architectures that were within our attentional framework, namely, CaffeNet [8], GoogLeNet [9] and VGGNet [10].

3 Methodologies

Our hybrid detection and localization methodology can be briefly outlined as follows: In a first feed-forward pass, a set of object class proposals is computed (Section 3.2) which are further analyzed via top-down backward propagation to obtain proposals regarding the location of the object in the scene (Section 3.2).

More specifically, for a given input image I, we begin by computing a set of object class proposals by performing a **feed-forward pass**. The probability scores for each class label (1 000 in total) are collected by accessing the network's

output *softmax* layer. Then, retaining our attention on the five highest predicted class labels, we compute the saliency map for each one of those predicted classes (see Figure 3). Then, a top-down **back-propagation pass** is done to calculate the score derivative of the specific class c. The computed gradient indicates which pixels are more relevant for the class score [11]. In the remainder of this section, we describe in detail the components of the proposed attentional framework.

3.1 Artificial Foveal Vision

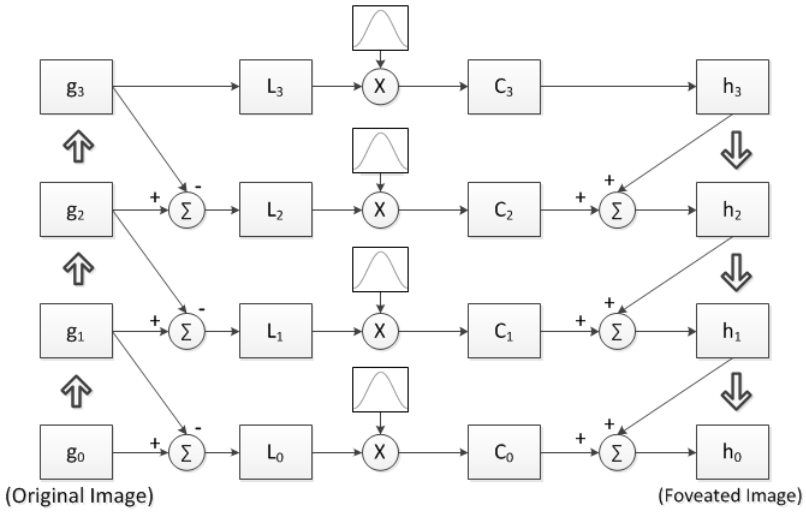
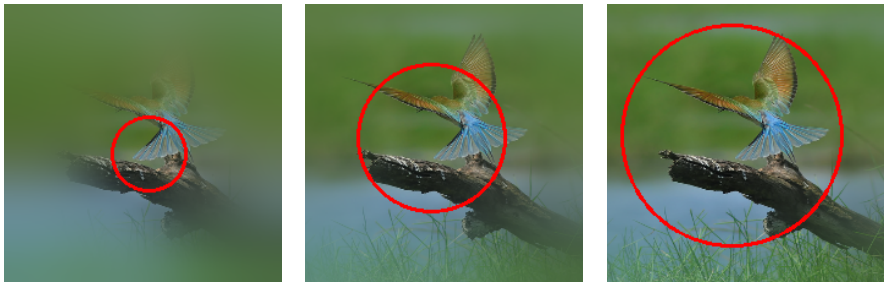


Fig. 1: A summary of the steps in the foveation system with four levels. The image g_0 corresponds to the original image and h_0 to the foveated image. The thick up arrows represent sub-sampling and the thick down arrows represent up-sampling.

Our foveation system is based on the method proposed in [12] for image compression (e.g. in encoding/decoding applications) which, unlike the methods based on log-polar transformations, is extremely fast and easy to implement, with demonstrated applicability in real-time image processing and pattern recognition tasks [13].

Our approach comprises four steps that go as follow. The first step consists on building a Gaussian pyramid with increasing levels of blur, but similar resolution. The first pyramid level (level 0) contains the original image g_0 that is down-sampled by a factor of two and low-pass filtered, yielding the image g_1 at level 1. More specifically, the image g_{k+1} can be obtained from the g_k via convolution



(a) $f_0 = 30$ (b) $f_0 = 60$ (c) $f_0 = 90$

Fig. 2: Example images obtained with our foveation system where f_0 defines the size of the region with highest acuity (the fovea), from a 224x224 image uniform resolution image.

with 2D isotropic and separable Gaussian filter kernels of the form

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

where u and v represent the image coordinates and $\sigma_k = 2^{k-1}\sigma_1$ the Gaussian standard deviation at the k-th level. These images are up-sampled to impose similar resolution at all levels. Next, we compute a Laplacian pyramid from the difference between adjacent Gaussian levels. The Laplacian pyramid comprises a set of error images where each level represents the difference between two levels of the previous output (see Figure 1). Finally, exponential weighting kernels are multiplied by each level of the Laplacian pyramid to emulate a smooth fovea. The exponential kernels are given by

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

where $f_k = 2^k f_0$ denotes the exponential kernel standard deviation at the k-th level. These kernels are centered at a given fixation point (u_0, v_0) which defines the focus of attention. Throughout the rest of this paper, we assume that $u_0 = v_0 = 0$. Figure 1 exemplifies the proposed foveation model with four levels and Figure 2 depicts examples of resulting foveated images.

Information Reduction The proposed foveal visual system is a result of a combination of low-pass Gaussian filtering and exponential spatial weighting. To be possible to establish a relationship between signal information compression and task performance, one must understand how the proposed foveation system reduces the image information depending on the method's parameters (i.e. fovea and image size).

Low-pass Gaussian Filtering Let us define the original high-resolution image as $i(u, v)$ to which corresponds the discrete time Fourier Transform $I(e^{jw_u}, e^{jw_v})$.

The filtered image $O(e^{jw_u}, e^{jw_v})$ at each pyramid level is given by the convolution theorem as follows

$$O(e^{jw_u}, e^{jw_v}) = I(e^{jw_u}, e^{jw_v}) * G(e^{jw_u}, e^{jw_v}). \tag{3}$$

Following the Parseval's theorem that describes the unitarity of a Fourier Transform, the signal information of the original image i is given by

$$E_i = \sum_{u=-\infty}^{+\infty} \sum_{v=-\infty}^{+\infty} |i(u,v)|^2 dudv = \frac{1}{4\pi^2} \int_{-\pi}^{\pi} \int_{-\pi}^{\pi} |I(e^{jw_u}, e^{jw_v})|^2 dw_u dw_v. \tag{4}$$

and the information in the filtered image o is given by

$$\begin{aligned} E_o &= \sum_{u=-\infty}^{+\infty} \sum_{v=-\infty}^{+\infty} |o(u,v)|^2 dudv \\ &= \frac{1}{4\pi^2} \int_{-\pi}^{\pi} \int_{-\pi}^{\pi} |I(e^{jw_u}, e^{jw_v}).G(e^{jw_u}, e^{jw_v})|^2 dw_u dw_v. \end{aligned} \tag{5}$$

Assuming that $I(e^{jw_u}, e^{jw_v})$ has energy/information equally distributed across all frequencies, of magnitude M, the information in the filtered image E_o can be expressed as

$$\begin{aligned} E_o &= \frac{M^2}{4\pi^2} \int_{-\pi}^{\pi} G(w_u)^2 dw_u \int_{-\pi}^{\pi} G(w_v)^2 dw_v \\ &= \frac{M^2}{4\pi^2} \int_{-\pi}^{\pi} e^{-w_u^2\sigma^2} dw_u \int_{-\pi}^{\pi} e^{-w_v^2\sigma^2} dw_v \\ &= \frac{M^2}{4\pi^2} \left(\frac{\pi \mathrm{erf}\left(\pi\sigma\right)^2}{\sigma^2} \right). \end{aligned} \tag{6}$$

Finally, the normalized information gain due to filtering for each level k of the pyramid is given by

$$P(\sigma_k) = \frac{1}{4\pi^2} \left(\frac{\pi \mathrm{erf}\left(\pi\sigma_k\right)^2}{\sigma_k^2} \right) \tag{7}$$

Gaussian Spatial Weighting The information due to spatial weighting is given by

$$R(f_k) = \frac{1}{N} \int_{-N/2}^{N/2} \int_{-N/2}^{N/2} e^{-\frac{1}{2}\frac{u^2+v^2}{f_k^2}} \, dudv \tag{8}$$

Hence, to compute the total information compression of the pyramid for the non-uniform foveal vision system, we need to take into account the combined information due to filtering and spatial weighting at each level of the pyramid. The total information of the pyramid is thus given by

$$T(k) = \sum_{k=0}^{K-1} R(f_k) L_k \tag{9}$$

where

$$L_k = P_k - P_{k+1} \quad \text{with} \quad P_0 = 1 \tag{10}$$

3.2 Weakly Supervised Object Localization

In this subsection we describe in detail our top-down object localization via feedback saliency extraction.

Image-Specific Class Saliency Extraction As opposed to Itti's [14] that processes the image with different filters to generate specific feature maps, Cao [2] proposed a way to compute a saliency map, in a top-down manner, given an image I and a class c. The class score of an object class c in an image I, $S_c(I)$, is the output of the neural network for class c. An approximation of the neural network class score with the first-order Taylor expansion [2][11] in the neighborhood of I can be done as follows

$$S_c(I) \approx G_c^\top I + b \tag{11}$$

where b is the bias of the model and G_c is the gradient of S_c with respect to I:

$$G_c = \frac{\partial S_c}{\partial I}. \tag{12}$$

Accordingly, the saliency map is computed for a class c by calculating the score derivative of that specific class employing a **back-propagation pass**. In order to get the saliency value for each pixel (u, v) and since the images used are multi-channel (RGB - three color channels), we rearrange the elements of the vector G_c by taking the maximum magnitude of it over all color channels. This method for saliency map computation is extremely simple and fast since only a back propagation pass is necessary. Simonyan *et al.* [11] shows that the magnitude of the gradient G_c expresses which pixels contribute more to the class score. Consequently, it is expected that these pixels can give us the localization of the object pertaining to that class, in the image.

Bounding Box Object Localization Considering Simonyan's findings [11], the class saliency maps hold the object localization of the correspondent class in a given image. Surprisingly and despite being trained on image labels only, the saliency maps can be used on localization tasks.

Our object localization method based on saliency maps goes as follow. Given an image I and the corresponding class saliency map M_c, a segmentation mask is computed by selecting the pixels with the saliency higher than a certain threshold, th, and set the rest of the pixels to zero.
Considering the stain of points resulting from the segmentation mask, for a given threshold, we are able to define a bounding box covering all the non-zero saliency pixels, obtaining a guess of the localization of the object (see Figure 3).

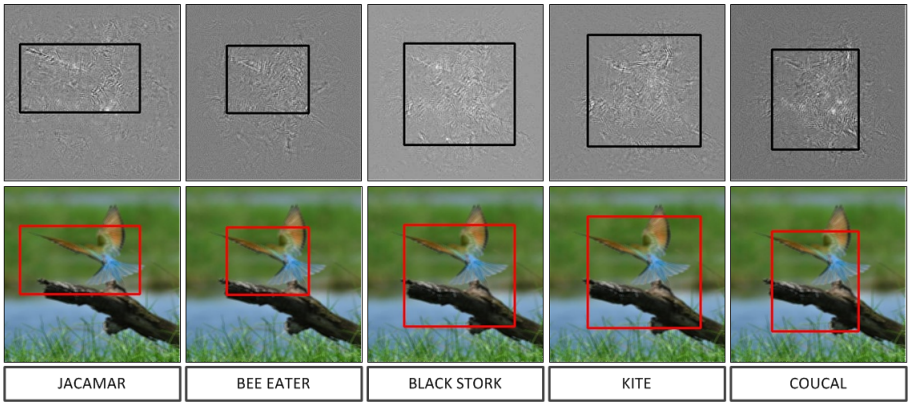


Fig. 3: Representation of the saliency map and the correspondent bounding box of each of the top 5 predicted class labels of a *bee eater* image of ILSVRC 2012 data set. The rectangles represent the bounding boxes that cover all non-zero saliency pixels resultant from a segmentation mask with $th = 0.75$.

4 Results

In this section, we begin by numerically quantifying the proposed non-uniform foveation mechanism information compression dependence on the fovea size. Then, we quantitatively assess the classification and localization performance obtained for the proposed feed-forward and feed-backward passes for various state-of-the-art CNN architectures (section 4.2).

4.1 Information Compression

In order to quantitatively assess the performance of our methodology, it is important to first quantify the amount of information preserved by the proposed non-uniform foveation mechanism to further understand the fovea size influence in task performance. Through a formal mathematical analysis of the information compression (see section 3.1) we can represent the relationship between fovea size (f_0), image size (N) and information compression. In our experiments σ_1 was set to 1, the original image resolution was set to $N \times N = 224 \times 224$ (the size of the considered CNNs input layers) and the size of the fovea was varied in the interval $f_0 = [0.1; 224]$. As depicted in Figure 4, the gain grows monotonically and exhibits a logarithmic behaviour for $f_0 \in [1; 100]$. Beyond $f_0 \approx 100$, the compression becomes residual, saturating at around $f_0 \approx 120$. Hence, from this point our foveation mechanism becomes unnecessary since resulting images contain practically the same information as the original uniform-resolution ones.

4.2 Attentional Framework Evaluation

In this paper, our main goal was to develop a single CNN capable of performing, recognition and localization tasks, taking into account both bottom-up and top-

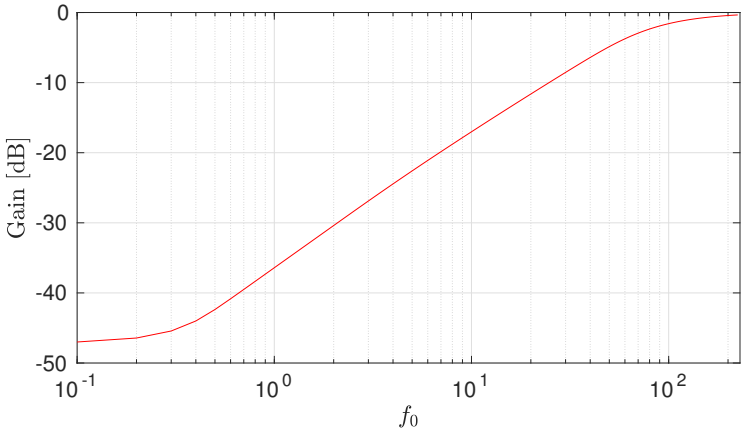


Fig. 4: Information gain in function of f_0 for the proposed non-uniform foveal vision mechanism.

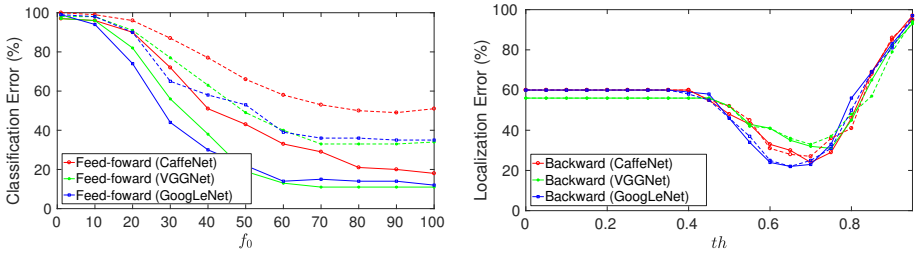
down mechanisms of selective visual attention. In order to quantitatively assess the performance of the proposed framework we used the ImageNet Large Scale Visual Recognition Challenge (ILSVRC) 2012[1]data set, which comprises a total of 50K test images with objects conveniently located in the images center.

Furthermore, we tested the performance of our methods with different pre-trained Convolutional Network (ConvNet) models which are publicly and readily available at Caffe [15] Model Zoo, namely, CaffeNet [8], GoogLeNet [9] and VGGNet [10]. As mentioned on Section 3.2, a feed-forward pass is executed originating a vector with the probability distribution of the class label scores. These class labels are used to compute the classification error which compares the ground truth class label provided in ILSVRC with the predicted class labels. Usually, two error rates are commonly used: the top-1 and the top-5. The former serves to verify if the predicted class label with the highest score is equal to the ground truth label. For the latter, we verify if the ground truth label is in the set of the five highest predicted class labels.

For a given image, the object location was considered correct if at least one of the five predicted bounding boxes overlapped over 50% with the ground truth bounding box. This evaluation criterion [16] consists on the intersection over union (IoU) between the computed and the ground truth bounding box.

Classification Performance The classification performance for the various CNN architectures combined with the proposed foveal sensing mechanism are depicted in Figure 5a. The CaffeNet pre-trained model which presents the shallower architecture had the worst classification performance. The main reason is that the GoogLeNet and VGG models use smaller convolutional filters and

[1] source: http://image-net.org/challenges/LSVRC/2012/ [as seen on June, 2017]



(a) Classification error (b) Localization error

Fig. 5: Classification and localization performance for various network architectures and sensing configurations. Left column: Dashed lines correspond to top-1 error and the solid ones correspond to top-5 error. Righ column: Dashed lines correspond to $f_0 = 80$ and solid lines to $f_0 = 100$.

deeper networks that enhance the distinction between similar and nearby objects.

Regarding the impact of non-uniform foveal vision, a common tendency can be seen for all three pre-trained models. The classification error saturates at approximately $f_0 = 70$. This result is corroborated by the evolution of the gain, depicted in Figure 4, since after -3 dB compressions goes slowly below 3 dB. This means that on average and for this particular data set, half of the information contained in uniform resolution images is irrelevant for correct classification.

Small size foveas exhibit extremely high error rates, which corresponds to a very small region characterized by having high acuity. This is due to the fact that images that make up the ILSVRC data set contain objects that occupy most of the image area, that is, although the image has a region with high-resolution, it may be small and not suffice to give an idea of the object in the image, which leads to poor classification performance.

Localization Performance As can be seen in Fig. 5b, for thresholds smaller than 0.4, the localization error remains consistent and stable at around 60%. From this point, the evolution of the error presents the form of a valley where the best localization results were obtained for thresholds between 0.65 and 0.7.

Overall, GoogLeNet presents the best localization performance. We hypothesize that this is mostly due to CaffeNet and VGG models feature two fully-connected layers of 4096 dimension that may jeopardize the spatial distinction of image characteristics. Furthermore, GoogLeNet is deeper than the aforementioned models and hence can learn discriminant features at higher levels of abstraction.

5 Conclusions and Future Work

In this paper we proposed a biologically inspired framework for object classification and localization that combines bottom-up ant top-down attentional mechanisms, incorporating recent Deep Convolutional Neural Networks architectures with human-like foveal vision. The main experimental goal of this study was to assess the performance of our framework with well-known state-of-the-art CNN architectures, in recognition and localization tasks, when combined with non-uniform foveal vision.

Through the analysis performed in our tests, we can conclude that the deeper neural networks present better performance when it comes to classification. Deeper networks have the capcity to learn more features which results in improved ability in distinguishing similar and close objects (i.e. generalization). Furthermore, the results obtained for non-uniform foveal vision are promising. We conclude that it is not necessary to store and transmit all the information present in a high-resolution images since, from a given fovea size (f_0), the performance in the classification task saturates.

However, the tests performed in this work assumed that the objects were centered, which is reasonable for the used data set, but unreasonable in real scenarios. In the future, we intend to test this type of vision in other data sets trained for recognition and localization tasks where objects may not be centered, thus having a greater localization variety. Dealing with multiple scales is another relevant limitation of non-uniform foveal sensors, in particular for close objects whose overall characteristics become unperceivable as the resolution decays very rapidly towards the periphery. To overcome this limitation, we intend to develop active vision mechanisms that will allow to autonomously redirect the attentional focus while integrating task-related evidence over time. Finally, it would also be interesting to train the system directly with blur (uniform and non-uniform foveal). In this case, it would be expected that with this tuning of the network, its performance would improve for both classification and localization tasks.

ACKNOWLEDGMENT

This work has been partially supported by the Portuguese Foundation for Science and Technology (FCT) project [UID/EEA/50009/2013]. Rui Figueiredo is funded by FCT PhD grant PD/BD/105779/2014.

References

[1] A Borji and L Itti, "State-of-the-art in visual attention modelling," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 35, no. 1, pp. 185–207, 2013.

[2] C. Cao, X. Liu, Y. Yang, Y. Yu, J. Wang, Z. Wang, L. Wang, C. Huang, T. S. Huang, W. Xu, D. Ramanan, and Y. Huang, "Look and Think Twice : Capturing Top-Down Visual Attention with Feedback," *IEEE International Conference on Computer Vision*, 2015.

[3] V. J. 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.

[4] R. S. Wallace, P.-W. Ong, B. B. Bederson, and E. L. Schwartz, "Space variant image processing," *International Journal of Computer Vision*, vol. 13, no. 1, pp. 71–90, 1994.

[5] Z Wang, *Rate scalable foveated image and video communications [ph. d. thesis]*, 2003.

[6] W. S. Geisler and J. S. Perry, "Real-time foveated multiresolution system for low-bandwidth video communication," in *Photonics West'98 Electronic Imaging*, International Society for Optics and Photonics, 1998, pp. 294–305.

[7] Y. LeCun, Y. Bengio, and G. Hinton, "Deep learning," *Nature*, vol. 521, no. 7553, pp. 436–444, 2015.

[8] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "ImageNet Classification with Deep Convolutional Neural Networks," *Advances In Neural Information Processing Systems*, pp. 1–9, 2012.

[9] C Szegedy, W Liu, Y Jia, and P Sermanet, "Going deeper with convolutions," *Computer Vision Foundation*, 2014.

[10] K. Simonyan and A. Zisserman, "Very Deep Convolutional Networks for Large-Scale Image Recognition," *International Conference on Learning Representations*, pp. 1–14, 2015.

[11] K. Simonyan, A. Vedaldi, and A. Zisserman, "Deep Inside Convolutional Networks: Visualising Image Classification Models and Saliency Maps," *Computer Vision and Pattern Recognition*, 2014.

[12] P. Burt and E. Adelson, "The laplacian pyramid as a compact image code," *IEEE Transactions on communications*, vol. 31, no. 4, pp. 532–540, 1983.

[13] M. J. Bastos, "Modeling human gaze patterns to improve visual search in autonomous systems," Master's thesis, Instituto Superior Técnico, 2016.

[14] 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, ISSN: 01628828. DOI: 10.1109/34.730558. arXiv: 0504378 [math].

[15] Y. Jia, E. Shelhamer, J. Donahue, S. Karayev, J. Long, R. Girshick, S. Guadarrama, and T. Darrell, "Caffe: Convolutional architecture for fast feature embedding," *arXiv preprint arXiv:1408.5093*, 2014.

[16] 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 (IJCV)*, vol. 115, no. 3, pp. 211–252, 2015.