Object detection and localization with Artificial Foveal Visual Attention

Cristina Melício, Rui Figueiredo, Ana Filipa Almeida, Alexandre Bernardino and José Santos-Victor
Institute for Systems and Robotics, Instituto Superior Técnico, Lisboa, Portugal
Email: cristina.melicio@tecnico.ulisboa.pt, ruifigueiredo@isr.tecnico.ulisboa.pt, ana.j.almeida@ist.utl.pt,
{alex, jasv}@isr.tecnico.ulisboa.pt

***Abstract*—In the last decades, in order to make the processing of a scene more efficient, biologically inspired approaches have been proposed. Visual attention models are being studied and actively developed in order to reduce the complexity and computational time of the existing methods. We propose a biologically inspired model that combines a single pre-trained CNN architecture with an artificial foveal visual system that performs simultaneously the classification and localization of objects in images. This model is based on the fact that only a small part of the image is processed with high resolution at each time so we load a foveated image in the network and successively employ feed-forward passes to determine the class labels and then via backward propagation determine the object possible locations according to each semantic label. By directing the attention to the center of the proposed location we mimic the human saccadic eye movements. In the results obtained we used the ILSVRC 2012 validation data set in a GoogLeNet CNN. We demonstrate that for non-centered objects the gain of the classification performance between iterations is significant showing that when mimicking the human visual behaviour of foveation, saccades are needed to integrate the information at each time.**

I. Introduction

The amount of information and visual stimuli that reaches the eyes is quite high, so the resources available by the human brain fail to process all of this visual perception information simultaneously [1]. For this reason, it is essential to process and interpret only the relevant stimuli. When an image is observed, the fixation region is projected onto the fovea and sampled with high density, while the periphery, on the other hand, is sampled at a lower resolution. There are also biological mechanisms that allow an active exploration [2] of the surroundings, namely the selective visual attention that gives priority to certain elements in a scene. These two aspects result in less information being processed by the brain at any given time but it requires the eyes to move constantly to integrate the information of the entire scene.

Likewise, computers also need to process information in real time, which requires a huge expenditure of resources. In the last decades, in order to make the processing of a scene more efficient, biologically inspired methods and approaches have been proposed [3]. These visual attention models are being studied and actively developed in order to reduce the complexity and computational time of the existing methods. Recent advances have been made in the field of artificial intelligence achieved by a new set of techniques known as Deep Learning. Among the techniques of Deep Learning are specially the Convolutional Neural Networks (CNNs). In face recognition or image classification with CNNs, the input image normally needs to be cropped so that objects are aligned roughly at the center of the image [4]. It could be advantageous in these cases to have an attentional model to select meaningful regions to process when the objects are not centered in the images. For this reason we propose an iterative foveation method that improves the classification of non-centered objects in a image.

In this work we study how to localize objects in a foveated image, where objects may lie on the periphery of the visual field. In these circumstances, the accuracy of the classification is low and a re-centering of the target is required to inspect the object with the high resolution part of the eye. Following the work done in [5], we consider the use of a classification CNN in a feed-forward and a feedback stages. In the feed-forward stage the network provides a ranked list of possible objects in the scene. In the feedback phase, these possible objects are located in the image in a top-down manner through the creation of a segmentation mask from the saliency map associated with the predicted class labels. Based on the foveation method proposed in [6] we considered in our work a foveal visual system to mimic the human visual information reduction and, in this way, the images loaded in the CNNs are foveated with different foveation points and different fovea sizes. Also, in visual search, humans tend to move the gaze towards objects, so our model is an iterative way of refining the classification and localization when performing each foveation step.

The main contributions of this paper are the following: first, we evaluate the performance of our methodology for a CNN architecture that can be used in tasks of detection and localization simultaneously when combined with human-inspired foveal vision. We also tried to understand the relationship between performance and different aspects: the fovea size, the segmentation mask threshold used in the localization, the foveation fixation point to mimic non centered objects and especially the gain in the performance between feed-forward passes.

The remainder of this paper is organized as follows: in section II we overview the related work and some fundamental concepts behind the proposed attentional framework. In section III we describe in detail the proposed methodologies, the artificial foveation system and a top-down, saliency-based mechanism for class-specific object localization. In section IV, we quantitatively evaluate our contributions. Finally, in section V, we wrap up with conclusions and ideas for future work.

II. Related Work

Visual perception arises when light is captured by the eyes and projected onto the retina. The human eye does not contain all the same visual acuity and the resolution of the captured image is much higher in the fovea, a small central region of the eye, decaying drastically as it approaches the periphery [7]. This non-uniform distribution leads to the need of moving the eyes towards the most important parts of the image in order to process them. There are also anatomical mechanisms for information selection, there are also functional mechanisms, such as attention, used to reduce the amount of information to be processed by higher cognitive levels of the brain. In the literature there are several types of computational foveation methods that attempt to replicate this human visual behavior: geometric [8], filtering-based [9] and multi-resolution methods [10].

A. Visual Attention

The concept of visual attention tries to explain how humans process the visual information that arrives to their eyes. The amount of visual information that is received by the eyes is quite high, about 10^8 to 10^9 bits per second [11] so it would be needed a high cognitive level and a great capacity of cerebral processing. Since brain resources are limited there are mechanisms to reduce the amount of information to be processed simultaneously [12].

Over the years, there have been several attempts to define visual attention. The most accepted definition solves the lack of cognitive resources and is called selective attention [13]. This concept consists of processing in more detail only subregions of the visual field, called focuses of attention, which are determined through selective mechanisms. According to the mechanism of selective attention the visual stimuli are ordered and processed in descending order of relevance, making the attention a sequential process. The most relevant stimuli are called salient. The relevance of visual stimuli can be influenced by the spatial location of the objects [14] and the a priori world knowledge; by certain features of objects present in the environment (color, size, orientation, direction of motion) regardless of their location [15]; or by the structure of certain objects [16].

B. Bottom-up and Top-down mechanisms

According to James [17] the selective orientation of attention to certain objects or locations is done through bottom-up and top-down factors. Bottom-up factors are driven by stimuli generated by features that are discriminative within a visual scene. Some features are intrinsically more salient in a given context, for example a black ball in the middle of white balls (the salient feature is the color). If a feature is visually salient from the surroundings, it automatically stands out and directs attention involuntarily. This suggests that the visual features are perceived in the brain before the attention itself [18]. On the other hand, top-down factors are generated by a goal to be performed and are influenced by knowledge, expectations and goals [19]. Attention driven by these factors is slower because it requires focal attention.

C. Deep Convolutional Neural Networks

In recent years have emerged new set of learning techniques known as Deep Learning [20]. These advances, were only achieved due to the development of more powerful hardware such as Graphics Processing Unit (GPUs) and the creation of a very large sets of labeled images (e.g. ImageNet [21]). Deep Convolutional Neural Networks (CNNs) are a class of deep artificial neural networks that are biologically inspired by the visual cortex of mammals. These networks have been widely used for image classification [22] and object detection [23]. These are based on a cascade of successive layers that apply different filters to the input data with the objective of extracting task specific features.

Some of the most used CNN architectures [24], [25] are inspired by the LeNet [26] network that follows the simple stacking structure of 7 convolutional layers interspersed with a *pooling* layer and then a last fully connected layer to perform the final classification. Similarly to the work of [5], we propose the use of a single pre-trained CNN, that combines covert (classification) and overt (localization) mechanisms of selective visual attention, with artificial foveal vision.

III. Methodologies

Our methodology, inspired by Cao's *et al* [5], combines a feed-forward classification with foveal selective mechanism and a feedback localization according to class labels set as goals in the visual search. We propose a biologically inspired foveal attention model that replicates the human visual system and it is capable of classifying and localizing objects in a image. This model is based on the fact that only a small part of the image is processed with high resolution at each time so we load a foveated image in the network and do a first feed-forward pass to determine the possible classes of objects in the image. Then via backward propagation we obtain several object locations according to each semantic label. By directing the attention to the center of the proposed location we foveate again the original image and re-classify and then we analyze by backward propagation the new locations of the new semantic labels.

We propose an iterative refinement model that improves the classification and localization. It can be decomposed into two phases: first the detection through an artificial foveation mechanism using the methodology of [6] and subsequent feed-forward classification, and second the localization by performing a back-propagation according to top-down information.

A. Artificial Foveal Visual System

In this work, we follow the foveation system proposed in [6] that tries to replicate the non-uniform distribution of the receptive fields in humans eyes. This artificial foveal system is inspired by the Laplacian Pyramid method proposed in [27]

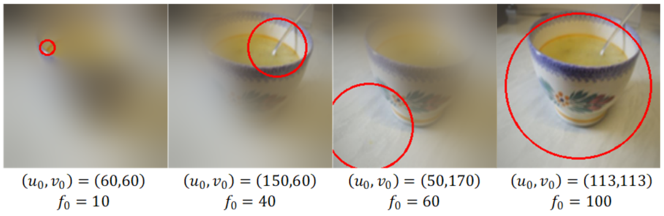


Fig. 1: Example of images with different foveal visual parameters as: foveation point (u_0, v_0) and fovea size f_0.

for image compression, which is extremely fast and easy to implement and has been applied in real-time image processing and pattern recognition.

This model consists of 4 steps:

1) First it is created a Gaussian pyramid where each level has increasing amount of blur and it is generated from the previous image level. Each subsequent image level is filtered using Gaussian kernel and scaled down. The image g_{k+1} can be obtained through the convolution of g_k with 2D isotropic and separable Gaussian filter kernels of the form

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

where u and v are the image coordinates, K is the number of levels of the pyramid and $\sigma_k = 2^{k-1}\sigma_1$ is the Gaussian standard deviation at the k-th level.

2) Secondly those g_k images are up-sampled to have the same resolution.

3) Then it is created a Laplacian pyramid by saving the difference between adjacent Gaussian level.

4) Finally, to mimic a high resolution in a f_0 size and a lower in the rest of the retina we multiply each level of the Laplacian pyramid by exponential kernels of the form

$$k(u, v, f_k) = e^{\frac{-(u-u_0)^2+(v-v_0)^2}{2f_k}} \qquad , 0 \le k < K. \qquad (2)$$

where $f_k = 2^k f_0$ is the exponential kernel standard deviation at the k-th level. The foveation point which defines the focus of attention is represented by (u_0, v_0).

In our work we want to vary this parameter to analyze the effect of non centered objects in an image. In figure 1 we represent different resulting images from our foveal visual system with different fovea sizes f_0 and foveation point (u_0, v_0).

B. Image-Specific Class Saliency Extraction

According to Simonyan's findings [28] it is possible to obtain an image-specific class saliency map via a back-propagation. Given an image I and a class c the CNN class score $S_c(I)$ is highly non linear therefore it is useful to approximate that with a first-order Taylor expansion in the neighborhood of I as

$$S_c(I) \approx G_c^{\intercal} I + b \qquad (3)$$

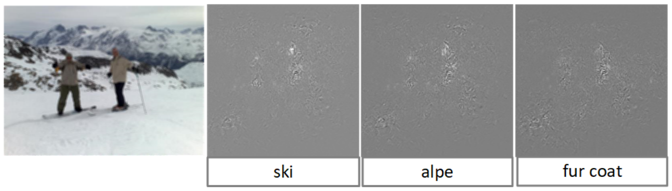


Fig. 2: Different saliency maps for specific class labels obtained by back-propagation in a top-down manner.

where b is the bias of the model and $G_c = \frac{\partial S_c(I)}{\partial I}$ can be viewed as a measure of how likely pixels of image I are important for the classification of a class c and therefore can give help us localize that class in the image. The pixel derivatives are found by back-propagation until the first input image layer. The back-propagated error values are the difference between the output of the CNN *softmax* layer and the desired output that corresponds to assign 1 to the input associated with the specific class we want to localize and assign 0 to all the other inputs. G_c defines the class specific salience map on image I. Since the images used are RGB a single class saliency value for each pixel $M_c(i, j)$ is obtained by taking the maximum magnitude of G_c across all colour channels l,

$$M_c(i, j) = \max_{l \in rgb} |G_c(i, j, l)|. \qquad (4)$$

C. Weakly Supervised Object Localization

The object localization is obtained by computing the segmentation mask by selecting the pixels of the saliency map M_c with a value higher than a certain threshold, th, and set the rest of the pixels to zero. A tightest bounding box covering the stain of non-zero saliency values is computed resulting in a guess of the localization of the object. Considering the center of the bounding box found we foveate again the original image and do the re-classification and re-localization of the image.

IV. Results

Following the work developed in [6] that considered only one feed-forward pass in the network with a centered foveated image, our main goal is to show that there is a significant gain in the performance between the first and the second foveation. Our model can be decomposed in the following steps (also illustrated in Fig. 3):

1) Resize the image to 227×227 and foveate with a specific fovea size f_0
2) Run CNN model with the foveated image and predict the top 5 class labels with a feed-forward pass
3) For each of the top 5 class labels, compute each localization bounding box with top-down back-propagation according to a threshold, θ
4) For each of the 5 bounding boxes in the original image foveate again with the fixation point in the center of each bounding box and predict again the top 5 class labels with a feed-forward pass
5) Given the total 25 labels and the corresponding confidences, rank them and choose the top 5 as final solution

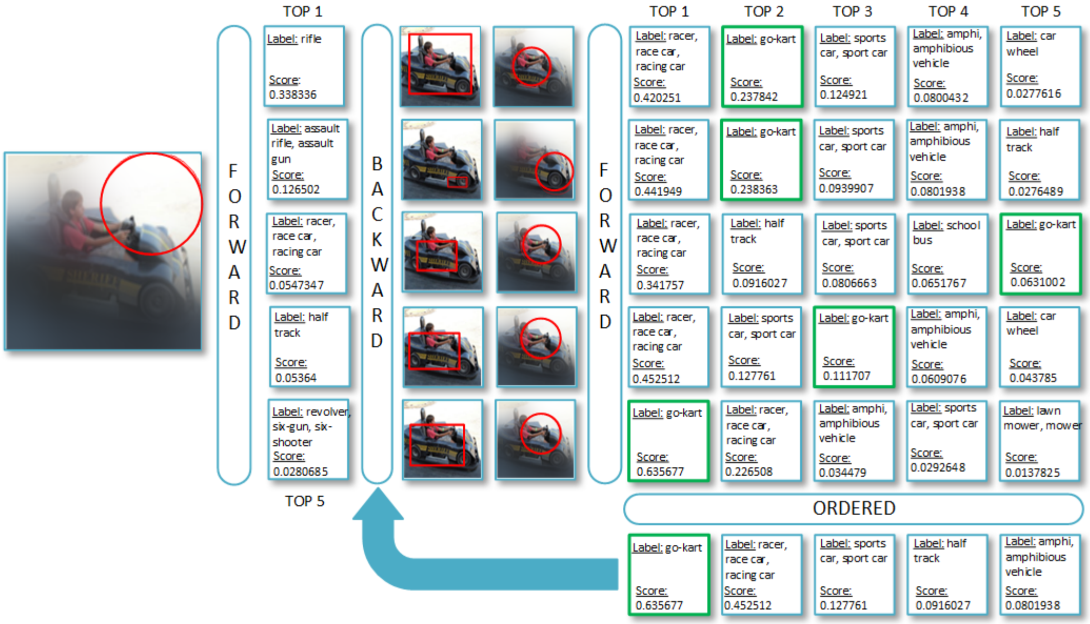

TOP 1
Label: rifle Score: 0.338336
Label: assault rifle, assault gun Score: 0.126502
Label: racer, race car, racing car Score: 0.0547347
Label: half track Score: 0.05364
Label: revolver, six-gun, six-shooter Score: 0.0280685
TOP 5
FORWARD
BACKWARD
FORWARD
TOP 1
TOP 2
TOP 3
TOP 4
TOP 5
Label: racer, race car, racing car Score: 0.420251
Label: go-kart Score: 0.237842
Label: sports car, sport car Score: 0.124921
Label: amphi, amphibious vehicle Score: 0.0800432
Label: car wheel Score: 0.0277616
Label: racer, race car, racing car Score: 0.441949
Label: go-kart Score: 0.238363
Label: sports car, sport car Score: 0.0939907
Label: amphi, amphibious vehicle Score: 0.0801938
Label: half track Score: 0.0276489
Label: racer, race car, racing car Score: 0.341757
Label: half track Score: 0.0916027
Label: sports car, sport car Score: 0.0806663
Label: school bus Score: 0.0651767
Label: go-kart Score: 0.0631002
Label: racer, race car, racing car Score: 0.452512
Label: sports car, sport car Score: 0.127761
Label: go-kart Score: 0.111707
Label: amphi, amphibious vehicle Score: 0.0609076
Label: car wheel Score: 0.043785
Label: go-kart Score: 0.635677
Label: racer, race car, racing car Score: 0.226508
Label: amphi, amphibious vehicle Score: 0.034479
Label: sports car, sport car Score: 0.0292648
Label: lawn mower, mower Score: 0.0137825
ORDERED
Label: go-kart Score: 0.635677
Label: racer, race car, racing car Score: 0.452512
Label: sports car, sport car Score: 0.127761
Label: half track Score: 0.0916027
Label: amphi, amphibious vehicle Score: 0.0801938

Fig. 3: Schematic of our iterative refinement model of object detection. First a foveated resized image is loaded into the network to predict the top-5 class labels through a feed-forward pass. Then for each class label, it is computed each bounding box with a top-down back-propagation according to the threshold selected. Then we apply a second foveation centered in each bounding box found and predict again the top 5 class labels with a feed-forward. Given this 25 labels with confidences associated we sort them in descending order, not choosing repeated labels and pick as final solution the top-5. Iteratively we to a re-localization according to those labels with a feedback pass. In our work we only considered two iterations. The red rectangles represent the bounding boxes that contain all pixels above the specified threshold, in this case the threshold was 0.75. The red circles represent the focused area simulating the fovea, that was set to $f_0 = 60$ in this case. The ground truth label of the input image is go-kart.

6) For each of the top 5 final class, compute the localization bounding box with top-down feedback pass

We used the ImageNet Large Scale Visual Recognition Challenge (ILSVRC) 2012 validation data set [29], which comprises a total of 50K test images with objects conveniently located in the images center. The following results were obtained with the first 100 images of that data set and the pre-trained neural network GoogLeNet [25].

According to our foveal visual system, in our experiments σ_1 was set to 1, the original image resolution was set to $N \times N = 227 \times 227$ (the size of the considered CNN input layer) and the size of the fovea was varied in the interval $f_0 \in \{0, \ldots, 180\}$. We could have considered 227 as upper limit, however, the size of the fovea becomes too large and there is no difference to the original image, and therefore, this fovea sizes larger than 180 not represent any benefit.

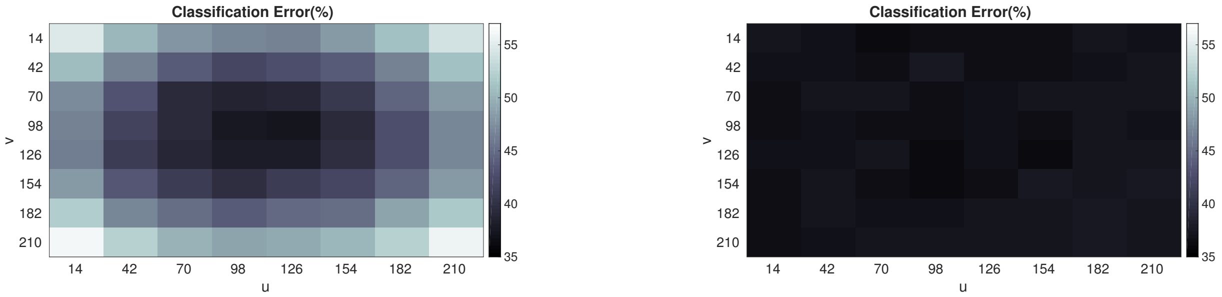
In order to quantitatively assess the performance of our methodology we considered the classification and localization error present in the ILSVRC [29].

A. Classification Performance

This classification performance is calculated for each image comparing the top-5 class labels in the descending order of confidence with the ground truth. If there is no match it leads to an error. The overall classification error is the average error over all images.

In order to understand how the foveation point of the first feed-forward pass influences the classification error, we made it vary along a 8 by 8 grid. As the threshold applied to the segmentation mask does not influence the classification error it was fixed to $\theta = 0.7$. However, the size of the fovea was varied between 0 and 180 so the classification error was calculated for each position over all f_0 considered. In Fig.4 we can compare the classification error between first and second feed-forward passes as a function of the foveation position. Since the objects of the data set are mainly centered, as we were expecting, the classification error is smaller in the center. However, we verified that from the first to the second pass, independently of the initial foveation point, the error reduces demonstrating the gain of our iterative model.

In order to understand better how the foveation size affects the classification error both for centered and non-centered foveation points we fixed $\theta = 0.7$ and varied f_0 between 0 and 180. In Fig. 5 we verify that the gain between the first and the second feed-forward classification is not significant when the foveation is centered, being at maximum 10%. However, when the foveation is non-centered (average of all foveation positions) the maximum gain between the two passes is 43%.



(a) First pass

(b) Second pass

Fig. 4: Classification performance in function of initial foveation point (u_0, v_0) where dark and bright represent better and worse performance. The classification error was calculated over all f_0 and fixing $\theta = 0.7$

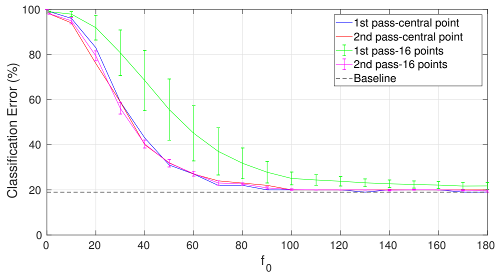


Fig. 5: Classification Performance in function of the fovea size f_0 with $\theta = 0.7$. The baseline was computed with $f_0 = 227$ (the resolution of the input image) to simulate an input image without any blur corresponding to minimum error

B. Localization Performance

The localization is considered correct if at least one of the five predicted bounding boxes for an image overlaps over 50% with the ground truth bounding box, otherwise the bounding box is considered wrong. The evaluation metric consists on the intersection over union between the proposed and the ground truth bounding box.

To understand the effect of the threshold, applied on the segmentation mask, on the localization performance, we fixed the fovea size to $f_0 = 70$ and varied the threshold in the interval $\theta = \in \{0, \ldots, 1.0\}$. As observed in Fig. 6(a) there is neither gain between backward passes nor differences between foveate in the center or elsewhere in the image. Localization tasks depend mostly on the low frequency of the image signal, thus, when we foveate an image we only remove high frequencies outside the fovea, however the location of the object remains detectable. For thresholds smaller than 0.4, the localization error remains stable. From this point, the evolution of the error presents the form of a valley obtaining the lowest localization error for thresholds of 0.65 and 0.7. This shows that exists a compromise between the threshold not being too small making nothing salient and being too high making all the image important for the bounding box. For this reason, we chose $\theta = 0.7$ to lead to minimum errors, when varying the fovea size as illustrated in Fig. 6(b).

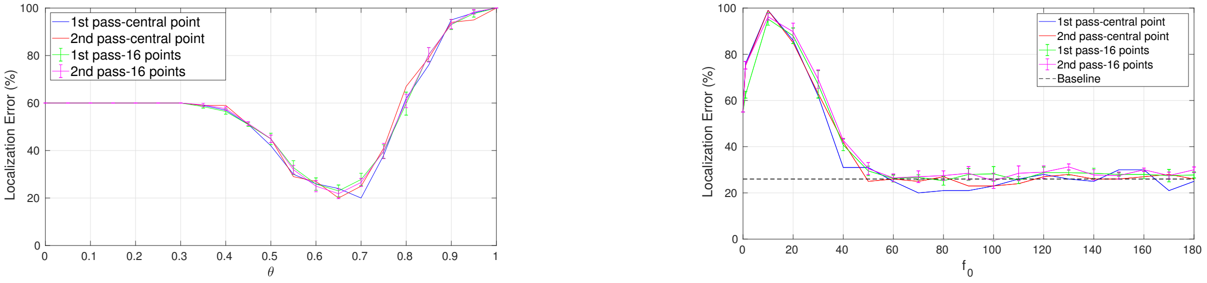
V. Conclusions

In this paper we proposed a biologically inspired framework for object classification and localization that incorporates CNNs with human-like foveal vision that mimics the selective attention mechanisms for information reduction to be processed by the brain. Our iterative model is composed by successive feed-forward and backward passes that refine the classification of objects.

The main experimental goal of this study was to assess the performance of our framework in tasks of detection and localization of non-centered objects in the images, to resemble real scenarios. The results obtained for our foveal iterative vision model are promising. We conclude, on one hand, that when using a methodology that replicates human visual behavior, it is necessary to use successive foveations (saccades). This is because in real scenarios, where objects can be anywhere in the image, the results show that the classification performance improves significantly from the first to the second feed-forward pass. On the other hand, we conclude that the classification performance reaches a saturation point for a fovea size of $f_0 = 70$. Furthermore, the localization performance does not improve with iterations. The location only depends on lower frequencies of the images and, thus, smoothing them with the foveation does not affect performance.

Our quantitative analysis indicates that for systems with foveation, which have a higher resolution in a small region that decays towards its periphery, we do not need the total resolution of the image to reach maximum performance. Thus, one can use mechanisms to reduce the resolution of an image, and it is not necessary to store and process all the information. However, we emphasize that the goal of this work was to study the impact of information reduction via space variant blurring of the original image, on classification and localization tasks using a state-of-the-art CNN classifier. Therefore, we did not show any computational gains, since the number of pixels of the input images were fixed.

In the future, in order to combine the mechanism of human-like visual saccades with computational gain, we intend to leverage log-polar like transformations or pyramidal images representations with more compact neural network architectures trained to classify images more efficiently.



(a)

(b)

Fig. 6: Localization Performance: (a) in function of the threshold applied to the segmentation mask with a fixed fovea size $f_0 = 70$; (b) in function of the fovea size f_0 where the threshold applied to the segmentation mask was set to $\theta = 0.7$ since it results in a minimium localization error.

Acknowledgements

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. Finally, we gratefully acknowledge the support of NVIDIA Corporation with the donation of the Titan Xp GPU used for this research.

References

[1] A. Borji and I. Laurent, "State-of-the-art in visual attention modeling," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 35, no. 1, pp. 185–207, Jan. 2013. [Online]. Available: http://dx.doi.org/10.1109/TPAMI.2012.89

[2] 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.

[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] T. Ngo and B. Manjunath, "Saccade Gaze Prediction using a Recurrent Neural Network," 2017. [Online]. Available: https://escholarship.org/uc/item/8qs6x5cv.pdf

[5] Y. Y. Y. Y. J. W. Z. W. Y. H. L. W. C. H. W. X. D. R. C. Cao, X. Liu and T. S. Huang, "Look and think twice: Capturing top-down visual attention with feedback convolutional neural networks," *IEEE International Conference on Computer Vision*, pp. 2956–2964, 2015.

[6] A. B. J. S.-V. A. F. Almeida, R. Figueiredo, "Deep networks for human visual attention: A hybrid model using foveal vision," *Third Iberian Robotics Conference*, 2017.

[7] A. L. Yarbus, "Eye movements and vision," *Neuropsychologia*, vol. 6, no. 4, pp. 389–390, 1968. [Online]. Available: http://linkinghub.elsevier.com/retrieve/pii/0028393268900122

[8] B. B. B.-E. L. S. Richard S. Wallace, Ping-Wen Ong, "Space variant image processing," *International Journal of Computer Vision*, vol. 13, no. 1, pp. 71–90, Sep 1994. [Online]. Available: https://doi.org/10.1007/BF01420796

[9] Laurent Itti, Christof Koch and E. Niebur, "A Model of Saliency-Based Visual Attention for Rapid Scene Analysis," *Short Papers*, vol. 20, no. 11, pp. 1254–1259, 1998.

[10] H. R. Sheikh, B. L. Evans, and A. C. Bovik, "Real-time foveation techniques for low bit rate video coding," *Real-Time Imaging*, vol. 9, no. 1, pp. 27–40, Feb. 2003. [Online]. Available: http://dx.doi.org/10.1016/S1077-2014(02)00116-X

[11] A. Borji and L. Itti, "State-of-the-art in Visual Attention Modeling," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 35, no. 1, pp. 185–207, 2013. [Online]. Available: http://ieeexplore.ieee.org/document/1307308/

[12] M. Carrasco, "Visual attention: The past 25 years," *Vision research*, vol. 51, no. 13, pp. 1484–1525, 2011.

[13] D. Heinke and G. W. Humphreys, "Computational models of visual selective attention: A review," *Connectionist models in cognitive psychology*, vol. 1, no. 4, pp. 273–312.

[14] M. I. Posner, "Orienting of attention." *The Quarterly Journal of Experimental Psychology*, vol. 32, pp. 3–25, 1980.

[15] A. Treisman and G. Gelade, "A Feature-Integration of Attention," *Cognitive Psychology*, vol. 136, pp. 97–136, 1980.

[16] J. Duncan, "Selective attention and the organization of visual information," *Journal of Experimental Psychology: General*, vol. 113, no. 4, pp. 501–517, 1984.

[17] G. A. M. William James, *The Principles of Psychology*. Harvard University Press, 1983, vol. Vols. 1-2. [Online]. Available: http://gen.lib.rus.ec/book/index.php?md5=A692A5A9B9632C6C605118CD493AE92F

[18] A. M. Treisman and G. Gelade, "A feature-integration theory of attention," *Cognitive psychology*, vol. 12, no. 1, pp. 97–136, 1980.

[19] M. Corbetta and G. L. Shulman, "Control of goal-directed and stimulus-driven attention in the brain," *Nature Reviews Neuroscience*, vol. 3, no. 3, pp. 201–215, 2002.

[20] I. Goodfellow, Y. Bengio, and A. Courville, *Deep Learning*. MIT Press, 2016, http://www.deeplearningbook.org.

[21] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, "ImageNet: A Large-Scale Hierarchical Image Database," in *CVPR09*, 2009.

[22] W. Rawat and Z. Wang, "Deep convolutional neural networks for image classification: A comprehensive review," *Neural computation*, vol. 29, no. 9, pp. 2352–2449, 2017.

[23] W. Zhiqiang and L. Jun, "A review of object detection based on convolutional neural network," in *2017 36th Chinese Control Conference (CCC)*, July 2017, pp. 11 104–11 109.

[24] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," in *Advances in neural information processing systems*, 2012, pp. 1097–1105.

[25] C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. E. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, and A. Rabinovich, "Going deeper with convolutions," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2015, pp. 1–9.

[26] Y. B. Yann LeCun, Léon Rottou and P. Haffner, "Gradient-Baseed Learning Applied to Document Recognition," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 1998.

[27] P. J. Burt and E. H. Adelson, "The laplacian pyramid as a compact image code," *IEEE Transaction on Communications*, vol. 31, pp. 532–540, 1983.

[28] A. Z. K. Simonyan, A. Vedaldi, "Deep inside convolutional networks: Visualising image classification models and saliency maps," 2014. [Online]. Available: https://arxiv.org/pdf/1312.6034

[29] 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.