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

Ana Filipa Almeida
Electrical and Computer Engineering, Instituto Superior Técnico, Lisboa, Portugal
Email: ana.j.almeida@ist.utl.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 incorporates bottom-up and top-down attentional mechanisms, combining 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 backward pass. At last, an image re-classification with attention is done by a second feed-forward pass. In this final stage, two visual sensing configurations are compared: a uniform (Cartesian) that uses a crop patch of the image to re-classify, discarding the surrounding context and a non-uniform tessellation that transforms the image by applying the human visual foveation model at the center of the object location proposal. The main contribution of our work lies in the evaluation of the performances obtained with uniform and non-uniform resolutions. We were able to establish the 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 task saturates.

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

I. Introduction

The available human brain computational resources are limited, so 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.

Humans use attention mechanisms based on goal-oriented (top-down) and stimulus-driven (bottom-up) information to define the region in the visual input where the attentional focus should be oriented [2]. In this way, the amount of processing is limited to a certain region of the visual field and the regions to explore (salient) are prioritized in time. Similar mechanisms can also be applied to artificial systems that share similar resource limitations. Much effort has been made towards understanding and applying the human attention mechanisms in robotic systems.

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 [3] which proposed to capture visual attention through feedback Deep Convolutional Neural Networks. Similarly in spirit, we propose a biologically inspired hybrid attention model, that combines bottom-up and top-down mechanisms and is capable of efficiently locate and recognize objects in digital images, using human-like vision.

More specifically, our method is constituted by three steps: first, we perform a feed-forward pass to obtain the predicted class labels. Second, a backward pass is made to create a saliency map that is used to obtain object location proposals after applying a segmentation mask. Finally, a second feed-forward pass is executed to re-classify the image with selective attention. With a non-uniform foveal visual sensor, the attention is directed to the proposed locations using a foveal spotlight model, whereas for the uniform sensor, the attentional spotlight is oriented in a covert manner to crop patches of the original image.

In this work, our primary goal is to evaluate the performance of several well-known Convolutional Neural Network architectures that are part of the state of the art in tasks of detection and localization of objects. Moreover, we assess the performance of two different visual sensory structures: a conventional uniform (Cartesian) and a multi-resolution, human-inspired, foveal configuration, on the first and second feed-forward passages. For the Cartesian, a re-classification is

performed for a cropped patch of the image, discarding the periphery. For the human-like sensor, the image is foveated at the center of the proposed location.

The remainder of this paper is organized as follows: Section II overviews the related work and some fundamental concepts needed for better understanding the proposed attentional framework. In Section III, a theoretical explanation of the saliency calculation for a specific object class is presented. The different sensing configurations are also presented, more specifically the uniform and the foveal vision topologies. The proposed hybrid attention model is presented on Section IV, where the various steps that constitute the framework are described. Finally, the obtained results are presented in Section V and in Section VI, we draw our conclusions.

II. Background

The proposed object localization and classification framework uses several biologically inspired attention mechanisms, which include space-variant vision, and Artificial Neural Networks (ANN) for top-down cognitive processes (i.e. guided, task-biased attention). 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.

A. Visual Attention

Attention is a process through which an organism selects a sub-region of the visual field, the so-called "focus of attention", to be processed in detail. This allows suppressing the rest of the available information to obtain an efficient perception. The observer attention can be stimuli-driven, triggered by scene characteristics like color or orientation (bottom-up factors) or by specific visual characteristics that depend on the task or goal that he wants to achieve (top-down factors).

B. Mechanisms for Information Processing

In general, when it comes to processing in the context of sensation and perception, two types of processing are commonly characterized: top-down processing and bottom-up processing. On one hand, **top-down** processing corresponds to allocate attention voluntarily on features, objects or spatial regions based on prior knowledge and current goals/tasks. Thus, prior knowledge and the task at hand are used to influence attention in a goal-driven manner. On the other hand, **bottom-up** processing refers to the involuntary mechanisms responsible for directing attention to salient regions based on differences from a region and its surround (e.g. contrast). In this case, the stimuli directly triggers our attention and, thus, it is a data-driven process.

C. Artificial Neural Networks

Artificial Neural Networks (ANN) are computational models inspired by the central nervous system of an animal, specially the brain, and try to mimic the way a biological brain solves problems. Its key element is the ability to learn implicit mappings between inputs and outputs, making it a powerful tool. They are also capable of recognizing patterns.

A neural network is organized in layers that establish connections between neurons. It starts with an input layer where each neuron is fully connected to all neurons at the next layer. To each connection between two neurons is assigned a weight that controls the signal transmission between them. The input units receive information from the outside world and communicate with one or more hidden layers where actual processing takes place. In classification networks, the hidden layers apply a kind of distortion of the input data in a non-linear way with the aim of having linearly separable categories at the end [4]. The last hidden layer links to the output layer where items are assigned to the most likely belonging class. All neurons in the hidden layers are processed by an activation function that can be linear, threshold or sigmoid function.

D. Convolutional Neural Networks

As far as visual attention is concerned, the most commonly used are the Convolutional Neural Networks (CNN), that are feed-forward ANNs that take into account the spatial structure. 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.

Convolutional layer: Each neuron receives a sub-region from a previous layer as input and these local inputs are multiplied by the weights. These filters are applied throughout input space with the purpose of looking for specific features. Their weights are shared and their output is a feature map.

Pooling layer: Its goal is to reduce the input dimensionality and produce a single output from the local region.

Fully-connected layer: Is the upper layer and computes the class scores to be consistent with training set labels. The input of the fully-connected layers is the set of all feature maps at the previous layer.

E. Deep Neural Networks

Deep neural networks are a subclass of 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 [4]. They used unsupervised learning methods to create layers of feature detectors without the need of labelled data. Then, they pre-trained some layers with more complex feature detectors providing enough information to initialize the weights with sensible values. This method allowed researchers to train networks 10 or 20 times faster [4]. In recent years, CNN are becoming deeper and deeper which resulted in a performance

boost. However, they are not becoming wider (number of parameters in each layer), since very wide and shallow networks exhibit very weak performance at generalization despite being strong at memorization. As opposed, deeper networks can learn features through several levels of abstraction and present much better results in generalization because they learn all the intermediate features between the raw data and the high-level classification. Note that using wider and deeper networks lead to an increase in the number of the parameters that the network will have to learn.

III. Hybrid Attention Model

In this section, we mention the saliency map concept and explain in detail how to compute the saliency map, in a top-down manner, for a given class. A uniform and a non-uniform foveal vision, two different sensing configurations are also presented.

A. Class Saliency Visualization

The need to locate objects quickly and efficiently gave rise to the method proposed by Itti [5], based on visual salience that proposed the most likely candidates and eliminates those that are less likely.
The visual features that contribute to the selection of attention of a stimulus (color, motion, orientation) are combined in a saliency map that has normalized information of the individual features maps. In order to get a saliency map, the input visual information is analyzed for visual neurons, sensitive to several visual features of the stimuli. This analysis is done in parallel through all visual field at multiple spatial and temporal scales, originating a series of feature maps where each map represents the amount of a certain visual resource at any place of the visual field. In each map, according to Koch and Ullman [6], a local saliency is determined for how different this location is from nearby locations in terms of color, orientation, motion, depth. The most salient location is a good candidate for attentional selection. Finally, all highlighted locations from all feature maps are combined in a single saliency map that represents a pure relevant signal which is independent of visual features.

As opposed to Itti's [5] method that computes the saliency map in a bottom-up manner, Cao [3] proposed a way to calculate the saliency map, in a top-down manner, given an image I and a class c. The class score $S_c(I)$ is a non-linear function of the image, hence an approximation of the neural network class score with the first-order Taylor expansion [3][7] in the neighborhood of I can be done as follows

$$S_c(I) \approx G_c^\top I + b \tag{1}$$

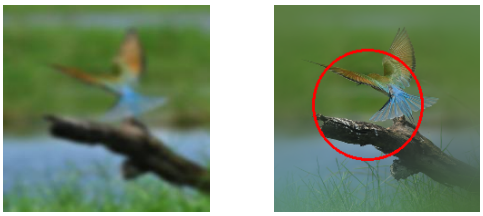
where b is the bias of the model and G_c is the gradient of S_c with respect to I:

$$G_c = \frac{\partial S_c}{\partial I}. \tag{2}$$

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 once 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.* [7] shows that the magnitude of the gradient G_c express 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.

B. Uniform vs Foveal Vision

In this work we will study and evaluate two types of organization of receptor fields: a conventional uniform distribution, typical in artificial vision systems (e.g. in standard image sensors), against a log-polar distribution, which approximates the human eye. The latter is composed by a region of high acuity – the fovea – and the periphery, where central and low-resolution peripheral vision occurs, respectively.



(a) $\sigma_0 = 5$ (b) $f_0 = 60$

Fig. 1: Two example images acquired with two different visual sensing configurations are shown: a uniform (a) with evenly blur of level σ_0 and a log-polar distribution (b) where f_0 defines the size of the region with high acuity.

1) Uniform Visual System: As many theories of visual processing proposed, a natural scene is processed in a fraction of a second [8] where a first rough description (the gist) of the scene is computed. Typically, imaging sensors use uniform resolution.

In the first feed-forward pass, we mimic the human behaviour on capturing the gist of the scene, quickly and with limited resources. For this matter, there is no need to rely on high-resolution images since this first glimpse takes only a split second and humans are capable of extract rough information of it [8]. In this way, we compress the images to save resources since in most cases, they are scarce.

For the initial glimpse, we want to simulate the use of a sensor with low-resolution, this is with lower level of detail which consequently requires fewer resources and comprehends a reduction of the information. However, image details correspond to edges that typically are only perceptible with high-resolution imaging sensors. For this purpose, the high-frequency details will be removed through low-pass filters.

When a low-pass filter is applied to a signal, its high-frequency components are completely removed. The simplest low-pass filter is the *ideal low-pass filter* that eliminates all

frequencies higher than a given *cut-off frequency* (f_c) and keeps the lower frequencies intact. Following this approach, we lose the high-frequency features like the edges. However, there is a way to remove the noise and preserve the edges and other (high-frequency) details. For this purpose, we use a Gaussian filter that does not abruptly remove high frequencies but soften them. The Gaussian filter alters the input image by convolution with an isotropic 2D Gaussian function that is defined as

$$G(u, v, \sigma) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{u^2+v^2}{2\sigma^2}} \quad (3)$$

where u and v represent the image coordinates and σ the standard deviation of the Gaussian distribution. The 2D Gaussian function is separable into u and v components thus we can perform first a convolution with a 1D Gaussian in the u direction, and then convolve with another 1D Gaussian in the v direction. In this study, we define σ_0 as the level of uniform blur (see Figure 1).

2) Foveal Visual System: 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 1). Many artificial foveation methods have been proposed in the literature that attempt to mimic similar behavior: geometric method [9], filtering-based method [10] and multi-resolution methods [11].

In this work, we rely on the method proposed in [12] for image compression (e.g. in encoding/decoding applications) which is extremely fast and easy to implement, with demonstrated applicability in real-time image processing and pattern recognition tasks as in [13]. This approach comprises four steps that go as follow. The first step consists on building a Gaussian pyramid. The first pyramid level (level 1) contains the original image g_1 that is low-pass filtered and down-sampled by a factor of two obtaining the image g_2 at level 2. The image g_3 can be obtained from the g_2 by applying the same operations, and so forth. The image g_{k+1} has a quarter of the resolution of image g_k where $k \in \{1, ..., K\}$ denotes the index of a pyramid level and K defines the total pyramid levels. This process is repeated as many times as the desired number of resolution levels for the pyramid.

In the next step, the Laplacian pyramid is build where the difference between the original image and the low-pass filtered image is computed. The Laplacian pyramid comprises a set of error images where each level represents the difference between two levels of the previous output (see Figure 2). Next, Gaussian weighting kernels are multiplied to each level of the Laplacian pyramid to implement the foveation mechanism. The Gaussian kernels are defined as in (3) and are generated just once and then displaced for a given point defining the focus of attention.

The next step consists of locating the foveation point which corresponds to the image location that will be displayed at the highest resolution. In our case, for the first feed-forward pass, the foveation point is defined as the center of the input image and for the second feed-forward pass, the foveation point is given by the center of the location proposal obtained through the analysis of the segmentation mask applied to the saliency map. At last, the foveated image is obtained by the reverse process used when building the Laplacian pyramid [12].

A summary of the human visual foveation model with four levels is presented on Figure 2. Starting with the original image, the levels g_1 to g_4 of the reduced pyramid are computed. Then, the difference between successive outputs from the previous step is obtained resulting the images L_1 to L_4 on the Laplacian pyramid. These images are multiplied by the kernels and an expand-and-sum procedure is done. An example of a foveated image obtained by this method is presented on Figure 1 where f_0 simulates the size of the fovea, central region of the retina of the human eye.

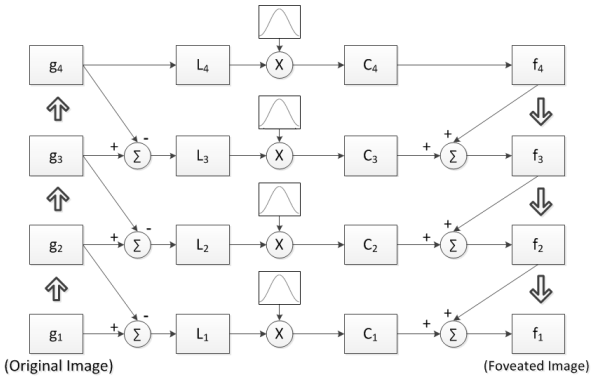


Fig. 2: A summary of the steps in the foveation system with four levels. The image g_1 corresponds to the original image and f_1 to the foveated image. The thick up arrows represent sub-sampling and the thick down arrows represent up-sampling.

C. Information Reduction

The different visual systems presented on Section III-B are based on different filtering strategies which result on reduction of information. To be possible to compare these systems, we have to understand how each system reduces the image information and what is the relationship between them.

1) Uniform Vision: The uniform visual system is computed via low-pass Gaussian filters. Let us define the original image as $i(u, v)$ to which corresponds the discrete time Fourier Transform $I(e^{jw_u}, e^{jw_v})$. The filtered image $O(e^{jw_u}, e^{jw_v})$ is given by the convolution theorem as follows

$$O(e^{jw_u}, e^{jw_v}) = I(e^{jw_u}, e^{jw_v}) * G(e^{jw_u}, e^{jw_v}). \quad (4)$$

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

$$E_i = \sum_{u=-\infty}^{+\infty} \sum_{v=-\infty}^{+\infty} |i(u, v)|^2 du dv = \frac{1}{4\pi^2} \int_{-\pi}^{\pi} \int_{-\pi}^{\pi} |I(e^{jw_u}, e^{jw_v})|^2 dw_u dw_v. \quad (5)$$

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

$$E_o = \frac{M^2}{4\pi^2} \int_{-\pi}^{\pi} G(w_u)^2 dw_u \int_{-\pi}^{\pi} G(w_v)^2 dw_v. \quad (6)$$

Furthermore, since we use $\sigma \geq 1$, the discrete time Fourier Transform is well approximated by the continuous time Fourier Transform. Thereby, we are now capable of simplify the expression of E_o presented on (6) as

$$E_o = \frac{M^2}{4\pi^2} \cdot \frac{\pi}{\sigma^2} = \frac{M^2}{4\pi\sigma^2}. \quad (7)$$

Finally, the information gain P is given by

$$P(\sigma) = \frac{E_o}{E_i} = \frac{1}{4\pi\sigma^2}. \quad (8)$$

For the uniform visual system, we apply to the image a uniform Gaussian blur for a given σ_0.

2) Non-Uniform Foveal Vision: For the non-uniform foveal vision, we implement the method explained on Section III-B2 where the blur is not evenly distributed, in the spatial domain. In the first step of our foveation system, we apply low-pass Gaussian filters and perform down-sampling in each level of the reduced pyramid.

The normalized information due to filtering for each level k of the pyramid is given by $P^k(\sigma_k)$ and the information due to spatial weighting is given by $R^k(f_k)$.

$$P^k(\sigma_k) = \frac{1}{4\pi\sigma_k^2}; \quad R^k(f_k) = \left(\frac{\int_{-N/2}^{N/2} e^{-\frac{1}{2}\frac{u^2}{f_k^2}} du}{N} \right)^2 \quad (9)$$

where N is the size of the image. Since the images are 2D, it is needed to calculate R^k for each dimension.

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

$$T(k) = \sum_{k=0}^{K} R^k P^k. \quad (10)$$

IV. Implementation

In this section, a detailed explanation of our model is made. In the first feed-forward pass, a rough description (the gist) of the scene is computed (Section IV-A) and analyzed via backward propagation to obtain proposals regarding the location of the object in the scene (Section IV-B). For the second feed-forward pass, two approaches have been compared, the human visual foveation model and the conventional uniform one (Section IV-C). For the former, an image re-classification is done directing the attention to the center of the proposed location. For the latter, the attention is directed to the cropped patch of the image, thus the remaining part of the image is discarded.

Our goal is to develop a single CNN capable of performing, at the same time, recognition and localization tasks taking into account both bottom-up and top-down mechanisms. For this purpose, creating a new data set seems unrealistic once plenty of images would have to be collected and hand labeled. Fortunately, a large visual data set of over 15 million labeled images called ImageNet is publicly available. For this work, we use the ImageNet Large Scale Visual Recognition Challenge (ILSVRC) 2012[1] data set, and some pre-trained Convolutional Network (ConvNet) models available at Caffe [14] Model Zoo. For this project the chosen pre-trained models were CaffeNet [15], GoogLeNet [16] and VGGNet [17] which present different depths.

A. Image-Specific Class Saliency Extraction

The pre-trained models CaffeNet, GoogLeNet and VGGNet were loaded to the corresponding networks for the test phase. Each network receives raw input data which needs to be pre-processed: subtract the mean over all images used in the training set in each color channel and swap color channels from RGB to BGR.

CaffeNet and GoogLeNet pre-trained models required a constant input dimension of 227×227 RGB images while VGGNet pre-trained model required a constant input dimension of 224×224 RGB images. Therefore the ImageNet images which present several resolutions were down-sampled for the required fixed resolution of the corresponding architecture. The ILSVRC 2012 validation set was used to perform the tests and evaluate our model.

After the pre-processing has been done, the network was loaded with images from the ILSVRC 2012 data set. The input images were transformed in two different ways: a uniform distribution and a foveal one as described on Section III-B1 and Section III-B2, respectively. We started by getting the network's output for the input image by performing a **feed-forward pass** filling the layers with data. Accessing the network's output layer of type *softmax*, the actual probability scores for each class label (1 000 in total) were collected.

Retaining our attention on the five highest predicted class labels which are more likely to be present in a given image, the saliency map for each one of those predicted classes was computed (see Figure 3). The method put into action to compute the saliency map was the one described on Section III-A where only an image I and a class c is required. As mentioned, a **back propagation pass** was done to calculate the score derivative of the specific class c. The calculation of the gradient tell us which pixels are more relevant for the class score [7].

B. Weakly Supervised Object Localization

Considering Simonyan's findings [7] mentioned on Section IV-A, the class saliency maps hold the object localization of the correspondent class in a given image. Surprisingly and

[1]source: http://image-net.org/challenges/LSVRC/2012/ [seen in November, 2016]

despite been 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 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).

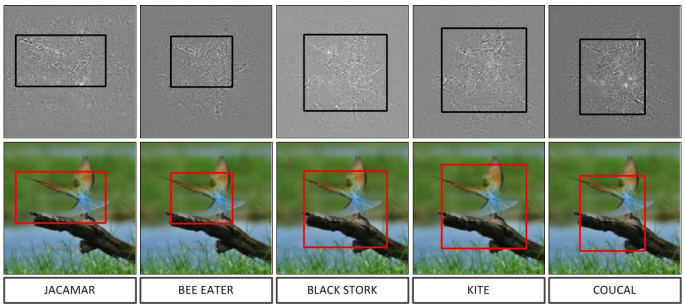


Fig. 3: Representation of the saliency map and the correspondent bounding box of each of the top 5 predicted class labels of a *bee eater* image of ILSVRC 2012 data set. The rectangles represent the bounding boxes that cover all non-zero saliency pixels resultant from a segmentation mask with $th = 0.75$.

C. Image Re-Classification with Attention

Given the initial guess of the object localization through the bounding boxes, the image labels are re-classified. We tested two different ways to re-classify image labels: firstly, inspired by Cao's work [3], we use cropped patches around the bounding boxes that are resized to the input dimension of the correspondent pre-trained model (227×227 for CaffeNet and GoogLeNet and 224×224 for VGGNet) and secondly, we foveate the images from the center of the bounding boxes with a fixed fovea size.

Following the first approach for image re-classification, the image patch was cropped from the original input image to ensure a good resolution and resized to the input dimension of the pre-trained model that supposedly corresponds to the smallest region that contained the object. Those new regions are then loaded into the neural network and a new **feed-forward pass** is done resulting in a re-classification of the regions. This strategy of re-classification is named by Cao [3] as the "*Look and Think Twice*" method.

For the second approach, there is no need to crop or resize the image. We use the bounding boxes obtained from the segmentation mask and apply the foveation method described on Section III-B. Considering that the bounding box provided by our framework contains the object, we direct our attention to the center of the bounding box and foveate the image for a given parameter f_0, highly specialized for high-resolution vision. The foveated image is then used as input to the network for the second feed-forward pass, giving rise to an image re-classification.

The image re-classification method (for both approaches) is applied to each of the five bounding boxes proposed from the first feed-forward pass where the highest five predicted class labels of each bounding box are preserved. Given the total 25 labels and the corresponding scores (confidence given by the network), we sort by descending order and pick the top-5 labels as the final solution. The sorted top-5 labels are then used to compute the classification error, corresponding to the second time we look to the image.

V. Results

In this section, the results and tests performed in this work are presented. We begin by establishing a numerical relationship between uniform and non-uniform visual systems on Section V-A. Next, an evaluation of the classification and localization performance obtained for the first and second feed-forward passes are done on Section V-B and Section V-C, respectively. Finally, on Section V-D the performance of the first pass is directly compared with the performance of the second pass for the different visual topologies.

A. Uniform vs Non-uniform Foveal Vision

Through the study of information gain done in Section III-C, we can represent the relationship between σ_0 and f_0, uniform and non-uniform vision, respectively (see Figure 4). With this analysis, it is possible to define the intersection point, that is, the values of σ_0 and f_0 where the information gain is the same for both types of sensors. Figure 4 was computed following the theory presented on Section III-C where expression (8) give us the evolution of the information gain for uniform vision and expression (10) for foveal vision.

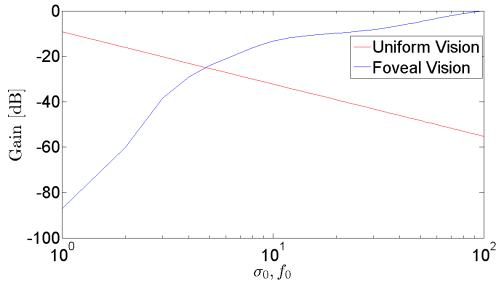


Fig. 4: Information gain in function of σ_0 for uniform and f_0 for non-uniform vision.

It is possible to verify that the information gain to the uniform vision is linear, in logarithmic scale, with respect to σ_0. As the blur level increases (σ_0), more information is compressed which leads to less gain. For non-uniform foveal vision there is a tendency to increase the information gain with the raise of f_0 but not linearly. This kind of evolution with f_0 for non-uniform vision makes sense since, as it increases, the size of the high-resolution region of the image also increases.

It is important to notice that for $f_0 = 100$, there is no gain of information, that is, for f_0 greater than 100, the processed image has the same information as the original one. The intersection point between the two different vision types is obtained with a gain of approximately -24 dB when $\sigma_0 = f_0 \approx 5$.

B. First Feed-Forward Pass

As mentioned on Section IV-A, 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 by ILSVRC with the predicted class labels. Usually, two error rates are commonly mentioned: 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 provided for the same image. If they are not a match, it leads to an error. For the latter, we verify if the ground truth label is in the set of the five highest predicted class labels.

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 a false positive [18]. The evaluation metric consists on the intersection over union between the proposed and the ground truth bounding box and this criteria was established on the ILSVRC 2012 challenge (see Figure 5).

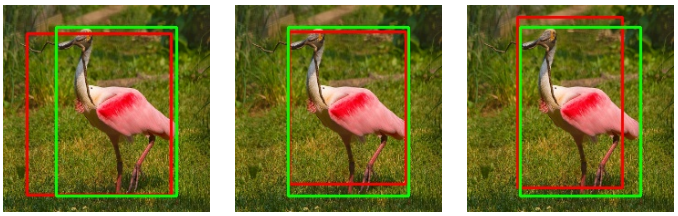


Fig. 5: We select a *spoonbill* image from the ILSVRC 2012 data set to demonstrate our weakly supervised object localization method. The red rectangles represent the computed bounding boxes for the top-3 predicted class labels and the greens, the ground truth bounding box of the *spoonbill* image.

The classification and localization errors were calculated for the three different topologies considered in this project: first, a non-uniform foveal vision where the images are foveated from the center for different f_0; second, a uniform vision characterized by having evenly blurred images for various σ_0 and finally, a combined vision that, for the first pass have uniformly blurred images with a level of blur equal to $\sigma_0 = 5$, which corresponds approximately to the point of intersection obtained from Figure 4.

1) Classification Performance: For the first pass, both with uniform and foveal sensors, we obtained the results presented on Figure 6. With regard to the classification, a global conclusion can be withdrawn: CaffeNet pre-trained model which presents the shallower architecture had the worst performance, obtaining the highest classification errors in all topologies. One possible justification for this is that the other GoogLeNet and VGG models use smaller convolutional filters and deeper networks that can enhance the distinction between similar and nearby objects.

For non-uniform foveal vision, a common tendency is visible on Figure 6a, in all three pre-trained models, there is a f_0 value from which the classification error saturates being this approximately $f_0 = 70$. This result is corroborated by the evolution of the gain depicted in Figure 4 where from $f_0 = 70$, the value of the gain is approximately -2 dB. This means that, beyond this fovea size, the amount of information that is added is not relevant for the correct classification of the object.

As expected for uniform (Cartesian) vision, as σ_0 increases and the blur level applied to the image rises, the amount of information present in the image decreases, resulting in an increase in the classification error. From Figure 6c it can be seen that this increase is approximately linear.

Through the relation obtained in Section V-A, we can compare the two types of vision, the uniform and the non-uniform foveal one. Thus, for $\sigma_0 = 5$, uniform vision presents a lower classification error, in the order of 50%. In turn, non-uniform foveal vision with $f_0 = 5$ shows an extremely high error. We hypothesize that the foveated area for this f_0 corresponds to a very small region characterized by having high acuity. The images that make up the ILSVRC data set have 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 performance in the classification task.

2) Localization Performance: The threshold parameter that defines which pixels will be selected to create the bounding boxes that represents the object location proposal. On one hand, if we set low thresholds, we will select all the pixels in the saliency map that have an intensity higher than this threshold, that is, we will base our localization task on a large number of pixels at the risk of having many outliers. On the other hand, the higher the threshold, the more restrictive the selection of pixels used for the localization. By visualizing the evolution of the location error as a function of the threshold, it is possible to verify that there is a trade-off between the chosen threshold and the location error obtained.

A consistent result in all topologies with respect to the localization error is the range of threshold values that get the smallest error. For thresholds smaller than 0.4, the localization error remains stable where the VGG model presents a smaller error compared to the other models. 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, depending on the sensor configuration used. GoogLeNet, the deeper model considered in this work, presents a better location performance compared to the other models in the range of thresholds located in the valley. Although the VGG model is deeper than the CaffeNet model, the latter has better

performance in the location. Both models feature two fully-connected layers of 4096 dimension that can ruin the spatial distinction of image characteristics. GoogLeNet does not have these fully-connected layers and has better results when it comes to the location of the object.

C. Top-Down Class Refinement

For the second pass, the topologies have undergone minor changes: first, in the non-uniform foveal vision, the foveation point ceases to be the center of the image and becomes the center of the bounding box proposals. Second, in the uniform (Cartesian) vision, instead of using the whole image for the re-classification, a cropped patch of the input image defined by the proposed bounding boxes is used, resulting in a loss of context but improved acuity. Finally, in the second pass of the combined vision, the original image is used with high-resolution and the foveal visual system presented on Section III-B2 is applied where the foveation point is given by the center of the bounding boxes.

The performance in the classification task in this second pass is cumulative, this is it depends on the parameters that were used in the first pass. Thus, the foveation point that in this second pass corresponds to the center of the bounding boxes, is dependent on the threshold that was used in the first pass and that gave rise to the location proposals. For the three topologies considered, the threshold chosen to be used in the segmentation mask and that conditions the location proposals in the second pass was $th = 0.7$.

For the uniform vision case, the input image used to be re-classified is the cropped patch of the original image defined by the location proposals. In this way, the presence of the context is discarded. Surprisingly, the presence or not of the context seems to have no great contribution in the classification of the object. One possible justification for this event is that each image contains only one object and this is found on a large scale. Thus, when the segmentation mask is applied to the saliency map, the proposed object location regions tend to contain the object itself (see Figure 5) and not so much information in relation to the context.

D. First vs Second Pass

For the non-uniform foveal vision, the difference between the first and the second pass is the selected foveation point in the input image: in the first, the image is foveated at the center and in the second, the foveation point is moved to the center of the proposed bounding boxes..

In Figure 6a and Figure 7a, it is possible to verify that there is practically no difference in classification error, this is, there is no significant difference in foveate from the center of the image or from the center of the object location proposal. One major limitation of this experiment is the fact that the objects are large scale and centered on the image. Therefore, we can conclude that for this data set and topology, there is no advantage in making a second pass through the network.

For the uniform vision, in the first pass, the image is evenly blurred for a given σ_0. In the second pass, a crop patch of the image defined by the proposed bounding boxes is used resulting in loss of context. As expected, regardless of the pass, the higher the blur level (σ_0), more information is reduced making it more difficult for the network to correctly classify the image, resulting in an increase of the classification error (see Figure 6c and Figure 7c).

The combined vision model is characterized for using images with a uniform blur of $\sigma_0 = 5$, in the first pass. In this case, the deeper the neural network used, the lower the classification error. This tendency remains in the second pass where predominantly, the deeper networks obtain better results. Again, due to the fact that the objects are centered in the image, as the size of the high-resolution region f_0 applied in the center of the location proposals increases, the lower classification error is obtained.

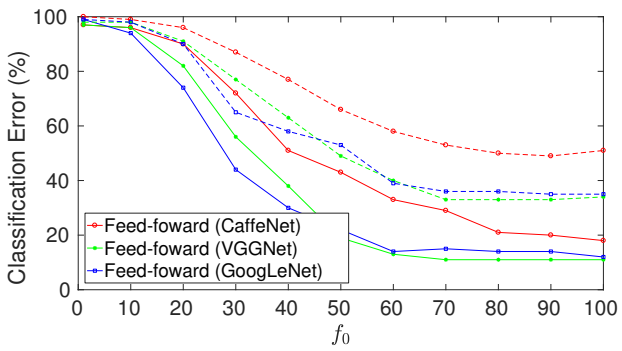
VI. Conclusions/Future Work

In this thesis we propose a biologically inspired framework for object classification and localization that incorporates bottom-up and top-down attentional mechanisms, combining the recent Deep Convolutional Neural Networks with foveal vision. We had as main goal of this study to evaluate the performance of several CNN architectures already known and usually used in recognition and localization tasks such as CaffeNet, VGGNet and GoogLeNet. Furthermore, we tested two different visual sensory structures, namely a uniform vision where it is not necessary to move the eyes towards the region of interest (covert attention) and a non-uniform foveal vision where the attention is directed to the location proposals of the object, by means of overt eye movements.

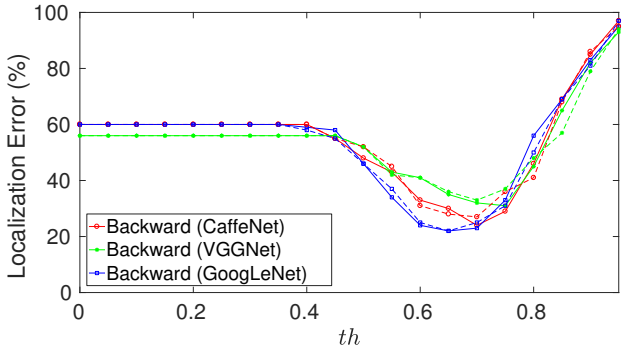
Through the analysis performed to our tests, we can conclude that the deep neural networks present better performance when it comes to classification. These deep networks have the ability to learn more features which results in a better learning in distinguishing similar and close objects.

The results we obtained for the 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 image since, from a given f_0, the performance in the classification task remains constant, regardless of the size of the high-resolution region.

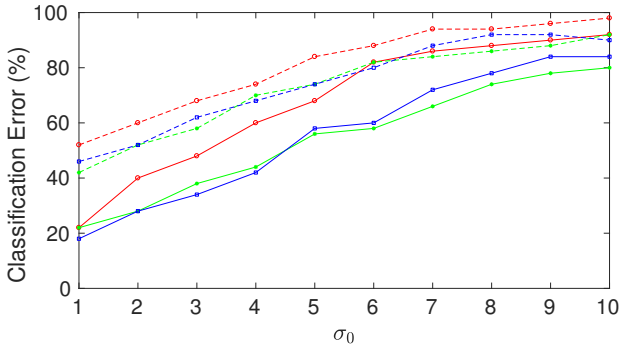
One of the major limitations on the evaluation of non-uniform foveal vision is that it is constrained by the chosen data set that presents the objects centered on the image. In the future, we intend to test this type of vision in other data sets trained for recognition and location tasks where objects are not centered, thus having a greater localization variety. The other very relevant limitation that also conditioned the tests is the scale of the images. Scaling is a problem for the foveal sensor in particular for very close objects because it loses the overall characteristics as the resolution decays very rapidly towards the periphery. 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 should improve for both classification and localization tasks.



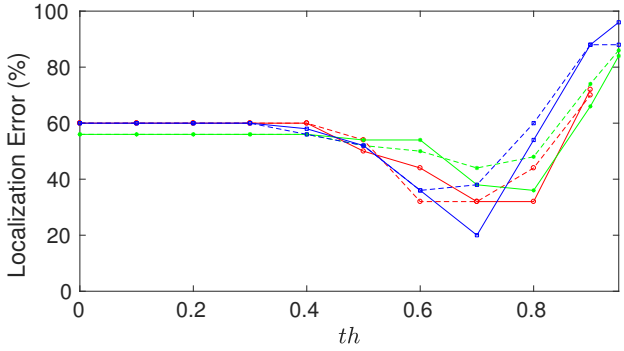
(a) Classification error: Non-uniform foveal vision.



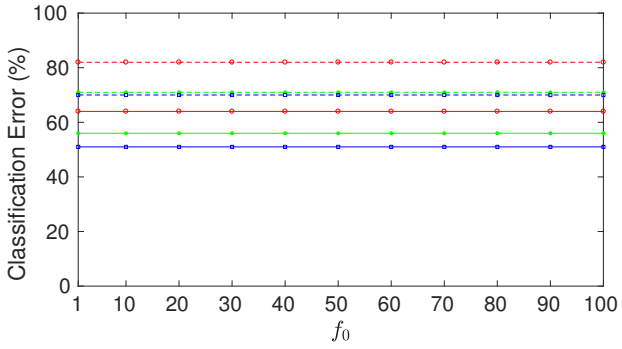
(b) Localization error: Non-uniform foveal vision.



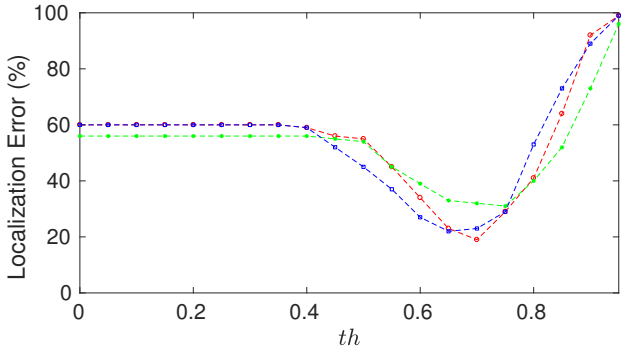
(c) Classification error: Uniform Cartesian vision.



(d) Localization error: Uniform Cartesian vision.



(e) Classification error: Combined vision.



(f) Localization error: Combined vision.

Fig. 6: Classification and localization performance of the **first pass** for various network architectures and sensing configurations. Left column correspond to classifcation error where dash lines correspond to top-1 error and the solid ones correspond to top-5 error.

References

[1] A Borji *et al.*, "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] F. Katsuki *et al.*, "Bottom-up and top-down attention: Different processes and overlapping neural systems," *The Neuroscientist*, vol. 20, no. 5, pp. 509–521, 2014.

[3] C. Cao *et al.*, "Look and Think Twice : Capturing Top-Down Visual Attention with Feedback," *IEEE International Conference on Computer Vision*, 2015.

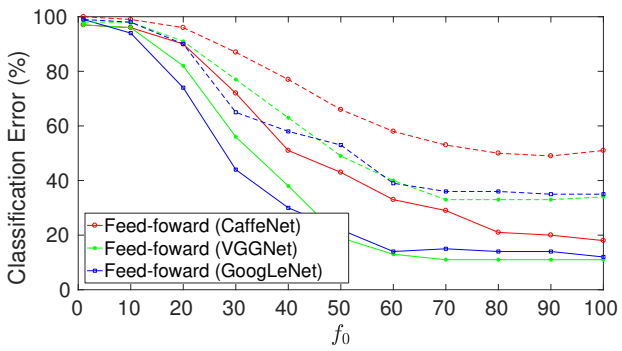
[4] Y. LeCun *et al.*, "Deep learning," *Nature*, vol. 521, no. 7553, pp. 436–444, 2015.

[5] L. Itti *et al.*, "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].

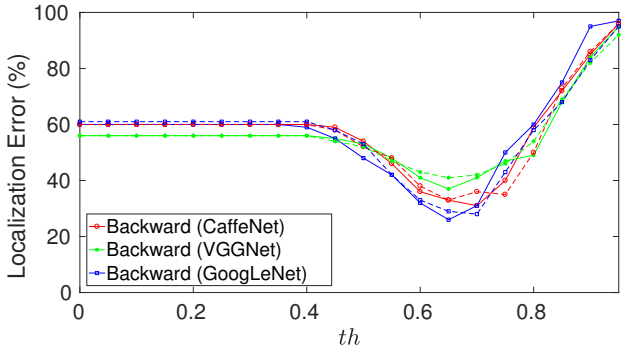
[6] C Koch *et al.*, *Shifts in selective visual attention: towards the underlying neural circuitry.* Springer Netherlands, 1985, pp. 219–27.

[7] K. Simonyan *et al.*, "Deep Inside Convolutional Networks: Visualising Image Classification Models and Saliency Maps," *Computer Vision and Pattern Recognition*, 2014.

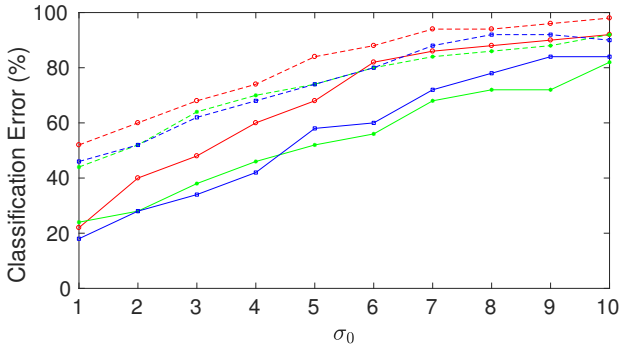
[8] I. I. A. Groen *et al.*, "From image statistics to scene gist: evoked neural activity reveals transition from low-level natural image structure to scene category.," *Journal of Neuroscience*, vol. 33, no. 48, pp. 18 814–18 824, 2013.



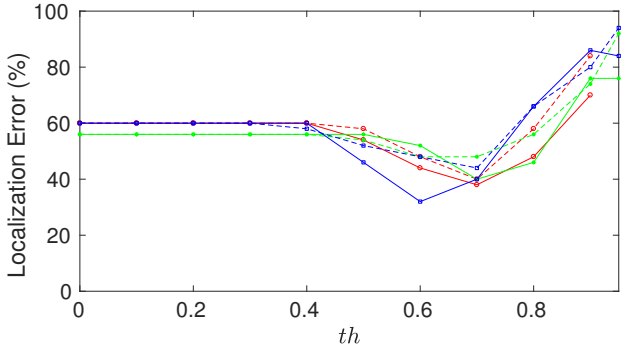
(a) Classification error: Non-uniform foveal vision.



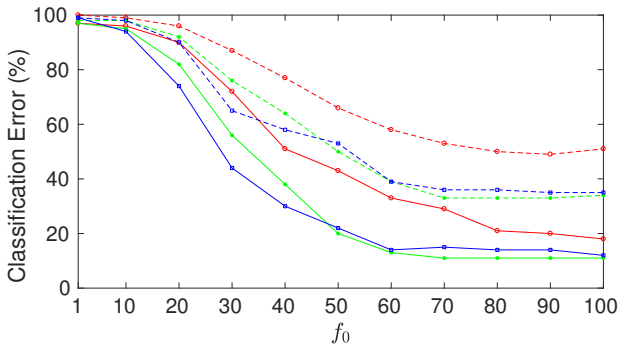
(b) Localization error: Non-uniform foveal vision.



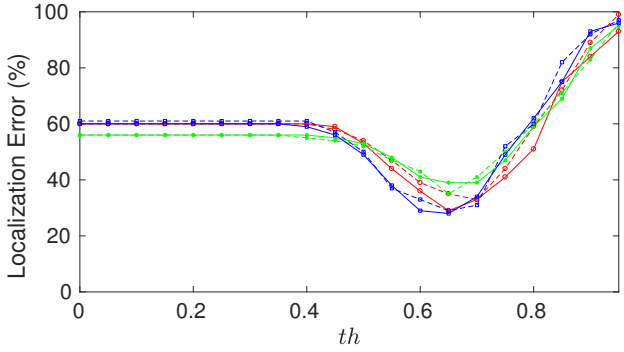
(c) Classification error: Uniform Cartesian vision.



(d) Localization error: Uniform Cartesian vision.



(e) Classification error: Combined vision.



(f) Localization error: Combined vision.

Fig. 7: Classification and localization performance of the **second pass** for several topologies. Left column correspond to classifcation error where dash lines correspond to top-1 error and the solid ones correspond to top-5 error.

[9] R. S. Wallace *et al.*, "Space variant image processing," *International Journal of Computer Vision*, vol. 13, no. 1, pp. 71–90, 1994.

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

[11] W. S. Geisler *et al.*, "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.

[12] P. Burt *et al.*, "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] Y. Jia *et al.*, "Caffe: Convolutional architecture for fast feature embedding," *arXiv preprint arXiv:1408.5093*, 2014.

[15] A. Krizhevsky *et al.*, "ImageNet Classification with Deep Convolutional Neural Networks," *Advances In Neural Information Processing Systems*, pp. 1–9, 2012.

[16] C Szegedy *et al.*, "Going deeper with convolutions," *Computer Vision Foundation*, 2014.

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

[18] O. Russakovsky *et al.*, "ImageNet Large Scale Visual Recognition Challenge," *International Journal of Computer Vision (IJCV)*, vol. 115, no. 3, pp. 211–252, 2015.