Retinotopy improves the categorisation and localisation of visual objects in CNNs

Jean-Nicolas Jérémie, Emmanuel Daucé, Laurent U Perrinet

Retinotopy improves the categorisation and localisation of visual objects in CNNs *

Jean-Nicolas Jérémie[1][0000−0002−9238−6654], Emmanuel Daucé[1,2][0000−0001−6596−8168], and Laurent U Perrinet[1][0000−0002−9536−010X]

[1] Aix-Marseille Université - CNRS
Marseille, France
[2] Ecole centrale Méditerranée, Marseille, France
jean-nicolas.jeremie@univ-amu.fr
emmanuel.dauce@univ-amu.fr
laurent.perrinet@univ-amu.fr

Abstract. Foveated vision is a trait shared by many animals, including humans, but its contribution to visual function compared to species lacking it is still under question. This study suggests that the retinotopic mapping which defines foveated vision may play a critical role in achieving efficient visual performance, notably for image categorisation and localisation. To test for this hypothesis, we transformed regular images by using a Log-polar mapping, and used this retinotopic images as the the input of convolutional neural networks (CNNs). We then applied transfer learning on pre-trained networks on the ImageNet challenge dataset. Our results show that surprisingly, the network re-trained on images which were compressed by the retinotopic mapping performs as well as the re-trained network applied to regular images. Moreover, we observed that the retinotopic mapping improves the robustness and localisation of image classification, especially for isolated objects. This was specially acute on a custom version of the dataset which aimed to categorise images that contain or not an animal. In summary, these results suggest that such retinotopic mapping may be an important component of preattentive processes, a central cognitive characteristic of more advanced visual systems.

Keywords: Foveated vision · Convolutional Neural Networks · Transfer learning · Visual categorisation · Neuromorphic transformation.

1 Introduction

The visual system in humans and many mammals is distinguished by a substantial resolution disparity between the central area of the visual field (fovea)

* Authors received funding from the ANR project number ANR-20-CE23-0021 ("AgileNeuroBot") and from the french government under the France 2030 investment plan, as part of the Initiative d'Excellence d'Aix-Marseille Université – A*MIDEX grant number AMX-21-RID-025 "Polychronies".

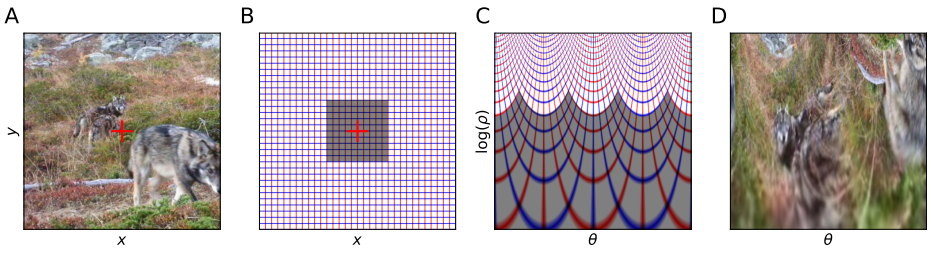

Fig. 1. We illustrate the process of transforming an example input image originally defined in Cartesian coordinates into retinotopic space using a Log-polar transformation. In **(A)**, the input image is presented with the fixation point marked by a red cross. The regular grid representing the image is defined by vertical (red) and horizontal (blue) Cartesian coordinates (x, y), as shown in **(B)**. As depicted in **(C)** to the image of the grid, by applying the Log-polar transformation, each pixel's coordinates with respect to the fixation point are converted based on its angle of azimuth θ (abscissa) and the logarithm of its eccentricity ρ (ordinates). This transformation results in an overrepresentation of the central area and a deformation of the visual space. When the transformation is applied to a natural image, as shown in **(D)**, there is a noticeable compression of information in the periphery.

and the peripheral regions, wherein the number of photoreceptors exponentially decreases with eccentricity [11]. Consequently, a natural question arises regarding the advantages conferred by these non-isotropic visual inputs in terms of information processing. Numerous hypotheses have been proposed regarding the role of this deformation of the visual field. One primary explanation is the coupling of foveal inputs with visual exploration : a retina with a fovea allows for efficient visual processing if the eye can actively move and focus its attention on specific points of interest. Studies have shown that this combination of saccades and foveal retina, coupled with an effective mechanism for detecting points of interest, significantly enhances visual acuity [4,5,3].

The most common approach to modeling foveal retinas involves reorganizing the pixels of an image into a Log-polar reference frame [9]. A Log-polar transformation organizes the visual field based on the angle and distance from the fixation point (eccentricity), with a resolution that exponentially decreases with the eccentricity. The primary role of a Log-polar transformation is to strongly compress the visual information, keeping high spatial frequencies at the center, but only low-spatial frequencies at the periphery. This conducts to process far less visual information when compared to the full resolution. Another important feature of the Log-polar transformation is the changing of the geometrical properties of the image, transforming rotations and zooms (homotheties) into translations [16].

We thus assess Log-polar visual processing on a well-known task, in the study of vision, that is the detection of an animal in a scene [6]. Applied to generic natural scenes, the task is such that the animal species is arbitrary. A

further difficulty is due to the large variations in identity, shape, pose, size, and position of the animals that could be present in the scene. Yet, biological visual systems are able to efficiently perform such detection in images which are briefly flashed [15]. Recently, deep learning algorithms have achieved an accuracy that is currently superior to humans for some visual recognition tasks. However, the tasks on which these artificial networks are typically trained and evaluated tend to be highly specialised and do not generalise well, e.g. accuracy drops after image rotation [8]. Here, we propose that a retinotopic mapping may be one essential ingredient in that robustness and study the advantages of this transformation in the context of image classification and localization.

2 Methods

2.1 Retinotopic mapping

We implement retinotopic mapping, as found in some animal species such as humans, so that visual information is concentrated at the center of gaze by applying a transformation from the regular Cartesian pixel grid to a Log-polar grid (see Figure 1). This transformation is accomplished using Pytorch library's [10] function : grid_sample(), it applies a grid to the pixels of the image in Cartesian coordinates. Therefore with a Log-polar grid, each pixel in Cartesian space is assigned a new position in Log-polar space. We set the number of angles sampled (N_θ) and the number of eccentricity sampled (N_ρ) to 256 to get an output image with a 256×256 resolution which was also used during the training process. All θ values are within a linear distribution in $[0; 2\pi]$, while ρ values are within a logarithmic distribution in $\log_2(r_{\min}; r_{\max})$. After analyzing various $r_{\min}$ parameters (performed with a central fixation point), we set $r_{\min}$ to -5; $r_{\max}$ fixes the radius and depend on the desired sub-sampling size. For instance, setting $r_{\max}$ to 0 gives maximal ρ values range within a $\log 2$ distribution in $[0.03; 1]$.

2.2 Transfer Learning

Transfer learning is a powerful technique that leverages knowledge gained from solving one problem, such as ImageNet [13], and applies it to a different yet related problem. Through our research, we successfully demonstrated the use of transfer learning to retrain VGG networks [14], enabling their application to various tasks. During the retraining process, we explored two network configurations: one with a retinotopic mapping at the input, and the other without. We have shown in our previous study that an appropriate training process is sufficient to produce performance with robustness comparable to physiological data [8]. Also we have shown that it is possible to predict the likelihood of a network trained on the animal task using the semantic link that connects the outputs of a pre-trained network to a label library such as ImageNet [8]. Therefore, we expect similar results even though we did not examine the networks re-trained on the animal task in this study. We extended the study by retraining

a Deep CNN ResNet101 on the categorization of 1000 ImageNet labels. This deeper network exhibited enhanced robustness, albeit at the cost of a higher computational load [7].

Each of these networks (i.e. Vgg16 and ResNet101) is then re-trained with Log-polar inputs and compared with the baseline network on the Imagenet dataset. Two types of task will be exploited: (i) categorization of a tag of interest among the 1000 labels in ImageNet and (ii) categorization and localization of an animal. The study covers 4 networks: Vgg16 Cartesian ImageNet and Vgg16 Polar ImageNet, ResNet101 Cartesian ImageNet and ResNet101 Polar ImageNet (where only Vgg16 Cartesian ImageNet and ResNet101 Cartesian ImageNet are not re-trained using transfer learning).

2.3 Data sets

We have selected two datasets for our study. The first dataset is ImageNet [13], which is widely used due to its extensive collection of images and associated labels. This dataset offers rich semantic links, enabling the construction of task-specific datasets, such as those focused on "animal" recognition. However, it is worth noting that ImageNet exhibits certain biases, particularly with objects being centered in many images. This characteristic makes it suitable for applying a Log-polar transformation, where information is concentrated around the fixation point, which is considered the center of the image during training.

Despite its advantages, ImageNet has limitations for localization tasks. For instance, it lacks multilabels, meaning there is only one label per image, and the proportion of bounding boxes relative to the image size is relatively small, which can limit the impact of certain analyses. To address these limitations, we also utilize the Animal 10k [17] dataset. This dataset provides key points for each animal present in an image. By fitting Gaussians to these key points, we can generate heat maps centered around the label of interest, which, in this case, is 'animal', see Figure 2. This approach enables us to improve localization and better analyze the distribution of animals in the images.

2.4 Likelihood map protocol

The CNNs described above are designed to categorise images by providing a likelihood value for each label. This likelihood is a probability that is, a scalar between 0 and 1) which predicts the probability that the label is present in the image. This allows to take a binary decision ("presence" or not) by choosing the label corresponding to the top likelihood, for instance. In our setting, we can also take different views from a large image and compute the likelihood for each of these, allowing to compare which view provides the best likelihood ("Bootstraping"). Views may consist for instance of cropping sub-images centred on different fixation points, with the fixation points aligned on a regular grid in visual space, see Figure 3.

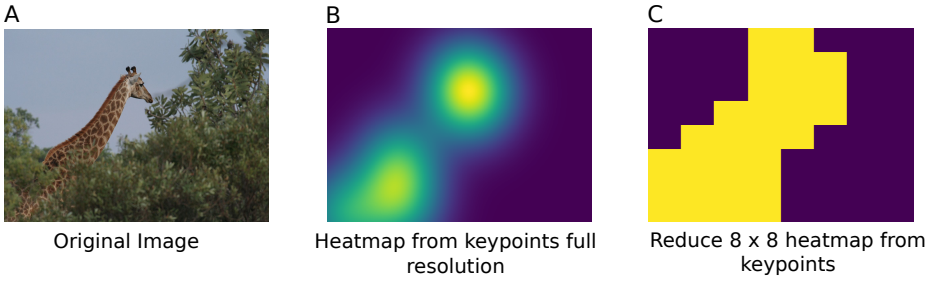

Fig. 2. (A) The original image of the Animal 10k dataset. (B) A heat map constructed by fitting Gaussians to the key points of the Animal 10k data set (see Methods : Data sets). (C) The heat map constructed in (B) is normalized and reduced to an 8×8 resolution to be used as ground truth when evaluating the heat map. A threshold (0.2) is applied to reduce the heatmap field to the assumed contour of the animal.

We used two parameters to define these maps: the first parameter is the resolution of the grid of fixation points. The second one is the size of the samples cropped at each of these positions define as the proportion of the input's Log-polar grid radius on the total input size (respectively Cartesian grid size, as the grid is a square for Cartesian samples see Figure 3-A & C). The input grid values determine the size of the sample taken. For a sample size of ratio 1.0 representing the entire input image, the grid values will lie within [-1.0;1.0], for a sample size of ratio 0.33 representing 30% of the total size of the input, the grid values will lie within [-0.33;0.33]. In the next section, we'll refer to the ratio of sample size to input size.

This sample is then transformed or not by the retinotopic mapping before being used as input for the corresponding network see Figure 3-B & D. Conveniently, a collection of samples for different fixation points can be process as a single batch, and we used here a range between 50 and 70 fixation points. This protocol define a likelihood map for any given network as the likelihood of categorising the presence of a label of interest (here "an animal") inferred at regularly spaced fixation points in the image.

3 Results

3.1 Average accuracy

We observed that the network retrained on transformed images had a similar categorisation accuracy to that of the network retrained on regular images. This is surprising, given that the networks were pretrained on regular images and that images with a Log-polar transformation show a high compression of visual information around the fixation point and a degradation of textures in the periphery, see Figure 4.

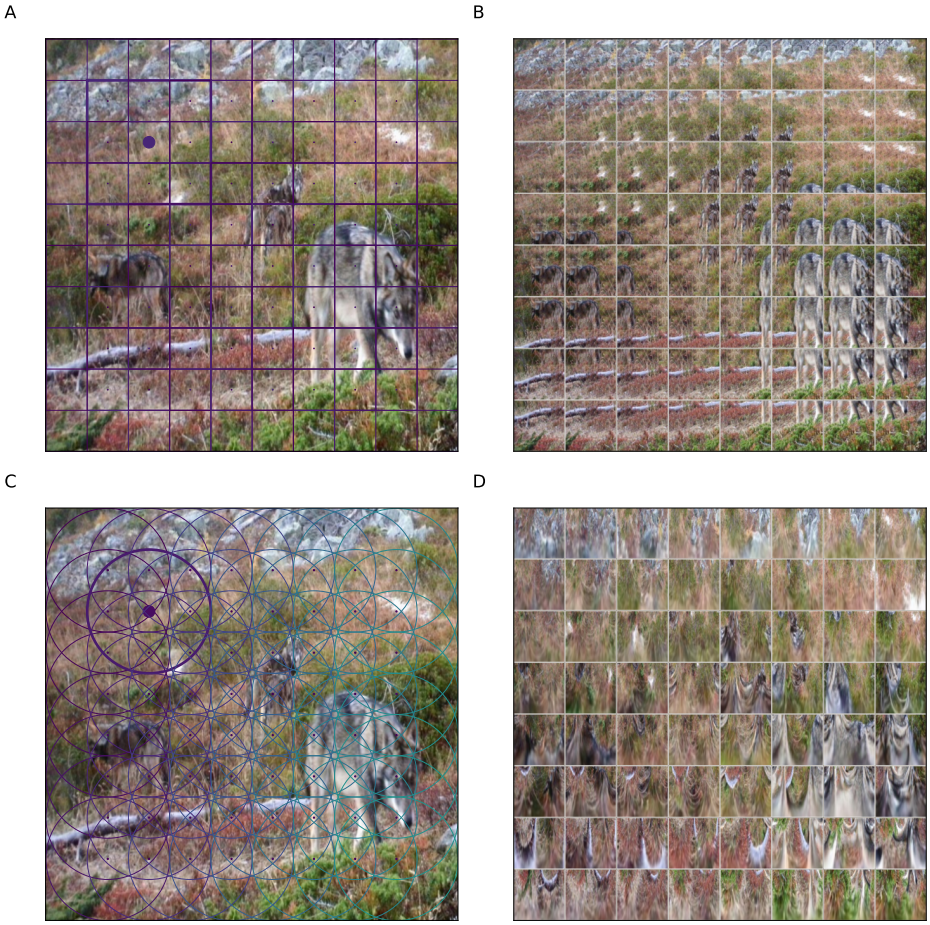

Fig. 3. Generating different views of a single image to compute likelihood maps. **(A)** For the networks using Cartesian inputs, we used a regular grid of 8×8 fixation points, which allow to crop samples, one particular view being highlighted. As shown in **(B)**, this creates a batch of images which can be used to generate likelihood maps. **(C)**) Similarly, we used a similar grid for generating batches of Log-polar inputs, as shown in **(D)**). In **(B)** & **(D)** each samples correspond to 33% of the input (see text for more details).

In addition, we found that while the VGG16 network retrained and tested on regular images showed some degradations for different rotations, the categorisation results were much more invariant for the network including a retinotopic mapping (see Figure 4). This phenomenon is a consequence of the translation invariance imposed by the structure of CNNs. Applied to the retinotopic mapping, this translation invariance in Log-polar space is transferred to a rotation and zoom invariance in the visual space [1]. The performance of RESNET101

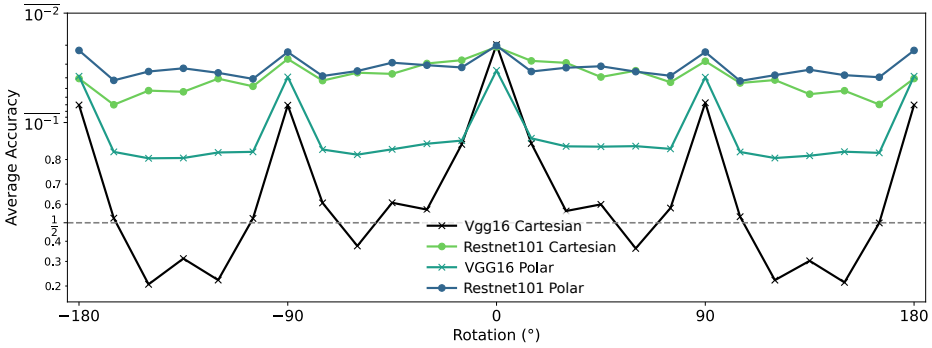

Fig. 4. Average accuracy over the ANIMAL 10K [17] dataset, shown for both retrained and pre-trained networks with different input image rotations. The rotation is applied around the fixation point with an angle ranging from $-180°$ to $+180°$ (in steps of $15°$). We tested each network (VGG16 or RESNET101) either with raw images or with retinotopic mapping (Cartesian or Polar). The dotted line represents chance level. This shows that VGG16 has a degraded performance compared to RESNET101, and notably that rotating images may have an adversarial effect on categorization performance, an effect which is less observed for RESNET101.

with Cartesian or Log-polar mapping are similar. Surprisingly, while this network was not designed a priori for retinotopic images, we observe a slight, but consistent, advantage for the retinotopic mapping.

3.2 Likelihood maps as a proxy for saliency

We tested the networks on the likelihood map protocol on a 8×8 fixed grid of fixation points varying the relative size of the input sample with different ratios (15%, 30%, 45%, 60%, see Table 1). Using the heat map extracted from the key points of the ANIMAL 10K [17] data set as ground truth, "in" represents coordinates inside an animal (and respectively "out" coordinates outside an animal, see Table 1). For each point in the 8 x 8 grid, a likelihood value is obtained (probability of an animal's presence). Next, we calculate the average likelihood for all points located within the zone corresponding to the animal (likelihood "in") as well as the average likelihood for the zone that does not contain the animal (likelihood "out"). Next, we compare the values obtained in the "in" zone with those obtained in the "out" zone. A higher contrast indicates the network's better ability to identify regions of interest in an image. For the RESNET101, both performed well on the task even if the CARTESIAN tend to maintain a high accuracy outside the box. For the RESNET101 networks, the CARTESIAN version of the network seems to perform much less well than the POLAR version in this exercise (see Table 1). If we consider a good categorization to be a high average probability on "in" coordinates (or a low probability on "out" coordinates), then in general, networks using POLAR grids tend to be slightly more contrasted than

networks using CARTESIAN grids, which is more manifest in the RESNET101 case. From this perspective, we observe that image ratios ranging between 30% and 45% appear to be best suited for highlighting the contrast between regions inside and outside the area of interest.

Table 1. Likelihood maps results for the VGG16 and RESNET101 networks and as computed on the IMAGENET challenge. Results are given as a fonction of the relative size of the samples with respect to the full image (Image Ratio). We highlight for each network the mapping which reaches maximal likelihood ratio for the "in" vs. "out" conditions.

	VGG16		RESNET101	
Ratio	Cartesian	Log-Polar	Cartesian	Log-Polar
15%	**1.18**	1.14	1.06	**1.14**
30%	1.19	**1.24**	1.06	**1.20**
45%	1.10	**1.19**	1.01	**1.14**
60%	1.03	**1.07**	1.01	**1.06**

3.3 Accuracy after "saccades" protocol

In this part of the study, we focused on finding a label of interest by including a large number of fixation points per image. Thus, in addition to the central fixation point (1 point with a sample ratio of 100%), we applied a grid of 7×7 fixation points (49 points with a sample ratio of 33%) as well as a grid of 3×3 fixation points (9 points with a sample ratio of 60%). All 59 fixation points are processed in a single image batch. The use of one of these fixation points would correspond to the network response after a saccade to an area of high salience.

We applied this protocol to the 50,000 images in the validation set of the IMAGENET data set. If we only stop at the best position (Top 1), the performance of the networks is degraded compared to their accuracy without saccades, and the same is true for Top 5 (compared to the performance of Top 5 without saccades, not shown here). On the other hand, by adding a simple saccade selection strategy (Top Choice), we find that the accuracy of all networks exceeds their baseline level.

4 Conclusion

A first and principal result of this study is proving the excellent capability of off-the-shelf Deep CNNs to deal with Log-polar inputs, that however represent a profound transformation of their visual inputs. The RESNET and VGG networks

Table 2. Accuracy after "saccades" results on the 1000 labels from ImageNet. Base represents the Top 1 accuracy of the network without saccades (state of the art accuracy), Top 1 represents the accuracy using the post-saccade maximum likelihood as the predictor, Top 5 represents the accuracy using the five post-saccade maximum likelihoods as the predictor. The Top choice represents the accuracy by taking the maximum post-saccade position if it is correct, otherwise we keep the pre-saccade prediction.

	Vgg16		ResNet101	
	Cartesian	Polar	Cartesian	Polar
Base	**0.74**	0.55	**0.78**	0.74
Top 1	**0.69**	0.55	0.67	**0.69**
Top 5	**0.79**	0.70	0.83	**0.84**
Top choice	**0.74**	0.64	**0.85**	0.80

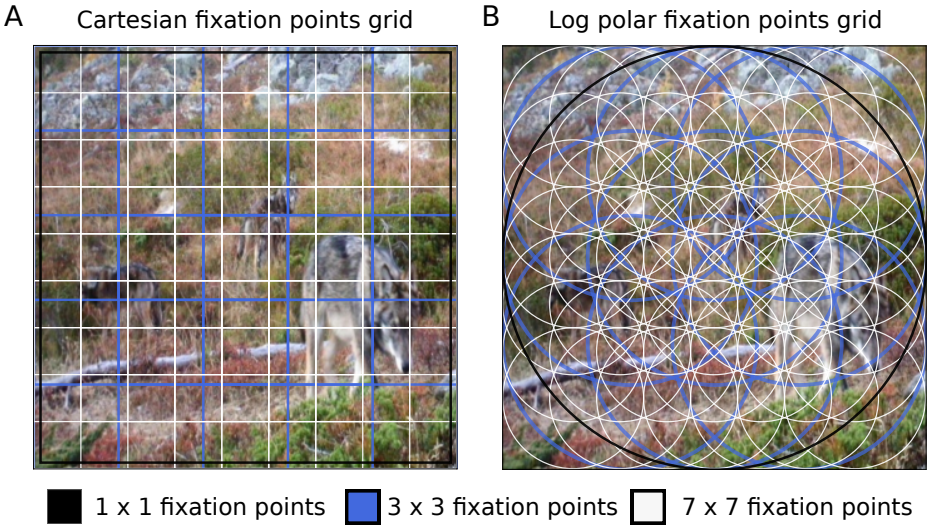

Fig. 5. (A) Example of a superposition of Cartesian fixation points (respectively Log-polar in **(B)**) used to carry out the after "saccades" protocol. With a central fixation point (black), a 3×3 grid of nine fixation points, each corresponding to a 60% ratio of the input (blue) and a 7×7 grid of forty-nine fixation points, each taking a sample corresponding to a 33% ratio of the input (white).

seem to effortlessly adapt to inputs where a large portion (the periphery) is heavily compressed, and the spatial arrangement significantly perturbed. The recognition rates achieved with Log-polar inputs are equivalent to those of the

original models. Additionally, the Log-polar transformation provides the added benefit of better invariance to zoom and rotation. However, this invariance comes at the expense of a reduced invariance to translation. For images that would not be centered on the region of interest, one would need to shift the fixation point to the area of interest, akin to eye saccades.

The integration of a retinotopic mapping approach holds significant promise for enhancing the efficiency and accuracy of image processing tasks. Our results are consistent with physiological data on ultra-rapid image categorisation [12,6]. The Log-polar compression employed in our approach allows for seamless extension to larger images without a significant increase in computational cost.

As a second result, the definition of saliency maps based on scanning the visual scene at a limited number of fixation points enables us to gain insights into Log-polar processing specificities: the Log-polar transformation provides a more focal view, thereby better separating the different elements of the image when focusing on its specific parts. In out case, it seems for instance to allow a more precise localisation of the category of interest, here an animal. It also gives us an insight into the features on which our networks actually rely. Such information can be compared with physiological data [2], used to design better CNNs, and ultimately allow physiological tests to be proposed to further explore the features needed to classify a label of interest. In particular, by focusing on the point of fixation with the highest probability in likelihood maps, we could envisage refining the training of the network our retinotopic mapping.

The accuracy performance of networks with a protocol that implements saccades in the process provides insight into the spatial modulation of network performance. It also allows us to extend the study of this type of network by implementing a strategy for choosing the optimal saccade.

Finally, the implementation of this robust categorisation, coupled with a refined localisation of a label of interest and the optimal selection of saccades, could allow us to extend this study to a more complex task. One such task is visual search (i.e., the simultaneous localisation and detection of a visual target), and the likelihood maps could provide the underlying pre-attentive mechanisms on which its effectiveness seems to depend.

References

1. Araujo, H., Dias, J.: An introduction to the log-polar mapping. Proceedings II Workshop on Cybernetic Vision (1), 139–144 (1997). https://doi.org/10.1109/CYBVIS.1996.629454, http://ieeexplore.ieee.org/document/629454/, 00000
2. Crouzet, S.M.: What are the visual features underlying rapid object recognition? Frontiers in Psychology **2** (2011)
3. Dabane, G., Perrinet, L.U., Daucé, E.: What you see is what you transform: Foveated spatial transformers as a bio-inspired attention mechanism. In: 2022 International Joint Conference on Neural Networks (IJCNN). pp. 1–8. IEEE (2022)
4. Daucé, E., Albiges, P., Perrinet, L.U.: A dual foveal-peripheral visual processing model implements efficient saccade selection. Journal of Vision **20**(8), 22–22 (Aug 2020). https://doi.org/10.1167/jov.20.8.22, https://jov.

arvojournals.org/article.aspx?articleid=2770680, 00003 Publisher: The Association for Research in Vision and Ophthalmology
5. Daucé, E., Perrinet, L.: Visual Search as Active Inference. In: Verbelen, T., Lanillos, P., Buckley, C.L., De Boom, C. (eds.) Active Inference. pp. 165–178. Communications in Computer and Information Science, Springer International Publishing, Cham (2020). https://doi.org/10.1007/978-3-030-64919-7_17, 00001
6. Fabre-Thorpe, M.: The characteristics and limits of rapid visual categorization. Frontiers in Psychology **2** (2011)
7. He, K., Zhang, X., Ren, S., Sun, J.: Deep Residual Learning for Image Recognition (Dec 2015). https://doi.org/10.1109/CVPR.2016.90, http://arxiv.org/abs/1512.03385, 336 citations (INSPIRE 2023/7/20) 336 citations w/o self (INSPIRE 2023/7/20) arXiv:1512.03385 [cs.CV]
8. Jérémie, J.N., Perrinet, L.U.: Ultrafast image categorization in biology and neural models. Vision **2** (2023)
9. Maiello, G., Chessa, M., Bex, P.J., Solari, F.: Near-optimal combination of disparity across a log-polar scaled visual field. PLoS computational biology **16**(4), e1007699 (2020)
10. Paszke, A., Gross, S., Massa, F., Lerer, A., Bradbury, J., Chanan, G., Killeen, T., Lin, Z., Gimelshein, N., Antiga, L., Desmaison, A., Kopf, A., Yang, E., DeVito, Z., Raison, M., Tejani, A., Chilamkurthy, S., Steiner, B., Fang, L., Bai, J., Chintala, S.: PyTorch: An imperative style, high-performance deep learning library. In: Wallach, H., Larochelle, H., Beygelzimer, A., dAlché-Buc, F., Fox, E., Garnett, R. (eds.) Advances in Neural Information Processing Systems 32, pp. 8024–8035. Curran Associates, Inc. (2019)
11. Polyak, S.L.: The retina. (1941)
12. Rousselet, G.A., Macé, M.J.M., Fabre-Thorpe, M.: Is it an animal? Is it a human face? Fast processing in upright and inverted natural scenes. Journal of Vision **3**, 440–455 (2003)
13. Russakovsky, O., Deng, J., Su, H., Krause, J., Satheesh, S., Ma, S., Huang, Z., Karpathy, A., Khosla, A., Bernstein, M., Berg, A.C., Fei-Fei, L.: ImageNet large scale visual recognition challenge. International Journal of Computer Vision (IJCV) **115**, 211–252 (2015)
14. Simonyan, K., Zisserman, A.: Very Deep Convolutional Networks for Large-Scale Image Recognition. arXiv:1409.1556 [cs] (2015)
15. Thorpe, S., Fize, D., Marlot, C.: Speed of processing in the human visual system. Nature **381**, 520–522 (1996)
16. Traver Roig, V.J., Bernardino, A.: A review of log-polar imaging for visual perception in robotics (2010)
17. Yu, H., Xu, Y., Zhang, J., Zhao, W., Guan, Z., Tao, D.: Ap-10k: A benchmark for animal pose estimation in the wild (2021)