Closed-loop visuomotor architecture supporting use of saccades for object recognition (D3.2 – SGA3)

Prototype cognitive architecture, integrating heterogeneous network modules and learning methods, with applications to visuo-motor related problems, including in-hand object manipulation.

Figure 1: The iCub robot in the Neurorobotics Platform (NRP) performing scene understanding.

Link to the experiment: https://github.com/ccnmaastricht/CDP4_NRP

Project Number:	945539	Project Title:	HBP SGA3

Document Title:	Closed-loop visuomotor architecture supporting use of saccades for object recognition
Document Filename:	D3.2 (D21) SGA3 M21 ACCEPTED 220520.docx
Deliverable Number:	SGA3 D3.2 (D21)
Deliverable Type:	Other
Dissemination Level:	PU = Public
Planned Delivery Date:	SGA3 M16 / 31 Jul 2021
Actual Delivery Date:	SGA3 M21 / 22 Dec 2021; accepted 20 May 2022
Author(s):	Mario SENDEN, UM (P117)
Compiled by:	Mario SENDEN, UM (P117) Vaishnavi NARAYANAN, UM (P117)
Contributor(s):	Mario SENDEN, UM (P117) Vaishnavi NARAYANAN, UM (P117) Alexander KRONER, UM (P117) Salil BHAT, UM (P117) Danny DA COSTA, UM (P117) Mahmoud AKL, TUM (P144) Anno KURTH, JUELICH (P20)
WP QC Review:	Yannick MOREL, UM (P117)
WP Leader / Deputy Leader Sign Off:	Rainer GOEBEL, UM (P117)
T7.4 QC Review:	N/A
Description in GA:	Closed-loop visuomotor architecture supporting use of saccades for object recognition. Prototype cognitive architecture, integrating heterogeneous network modules and learning methods, with applications to visuo-motor related problems, including in-hand object manipulation.
Abstract:	Human vision operates in light of physical constraints imposed by the eye as well as functional requirements such as having high visual acuity while maintaining a large field of view. Physically, the organisation of the retina meets these constraints by densely packing photoreceptors within a small central region, the fovea, and letting photoreceptor density decrease rapidly towards the periphery. Functionally, this arrangement introduces a challenge for the visual system since only the region of external space fixated by the eyes is resolved with high acuity. To overcome this, vision involves information integration across samples of external space obtained by continuously repositioning the eyes. As part of Deliverable 3.2, we have developed a closed-loop architecture consisting of five functional modules: retinal image sampling, saliency prediction, target selection, saccade generation, scene identification and embodied this in a virtual iCub robot simulated on the Neurorobotics Platform (NRP). The architecture engages in a free viewing paradigm, making saccades based on the saliency of regions in the visual scene, identifying fixated objects and identifying the scene (type of room) it finds itself in.

Keywords:	Scene understanding, saliency prediction, target selection, saccade generation, visuomotor functions, embodied cognition, neurorobotics, cognitive robotics, multiscale architecture, modular architecture, functional cognitive architecture
Target Users/Readers:	computational neuroscience community, neurorobotics community, computer scientists, consortium members, funders, general public, neuroscientific community, neuroscientists, platform users, policymakers, researchers, scientific community, students

Table of Contents

Table of Figures

1. Introduction

This work contributes towards developing a large-scale, modular, embodied architecture for visual scene understanding. To do so, this project brought together various modules implemented at various levels of abstraction using different modelling approaches. Using tools developed within the Human Brain Project, we demonstrate an architecture of saccades for scene understanding (SSU) loop. Scene understanding lends itself well to demonstrating the development of embodied architectures as it constitutes a high-level functional capacity that requires the coordinated interplay between numerous brain regions. Furthermore, scene understanding is a prime example of a functional capacity whose complexity can only be fully appreciated from an embodiment perspective.

Scene understanding is a visuomotor task which can be subdivided into scene recognition and saccade control. Scene recognition involves a hierarchical encoding process by which representations of (changing) light intensities are gradually transformed into a coherent percept. Saccadic control can be seen as a combination of three sub-capacities. The first is the computation of saliency distributions in a scene, a topographic map of the visual input highlighting distinctive regions with high information content. The next is target selection which involves a decision-making process to choose one of many salient regions for the next fixation. The last sub-capacity is saccade generation, triggering a ballistic eye movement toward the chosen target. In this work, individual capacities in this loop are implemented using different modelling approaches and different levels of abstractions. Specifically, we implement scene classification as well as saliency computation using *convolutional neural networks* (CNNs), target selection using generalised Lotka-Voltera neuronal population equations and the subcortical saccade generation circuit using leaky integrate-and-fire neurons.

The SSU architecture is embodied on the *Neurorobotics Platform* (NRP), developed in coordination with Service Category 4 (SC4). The NRP is a software infrastructure where experimentations can be carried out in a virtual environment. It allows connecting brain models to detailed simulations of robot bodies and environments. This is an invaluable resource for testing functional cognitive models in an embodied setting. Through collaboration with SC4, neuroscientists were provided with a platform to test their models, while the NRP developers gained insight into the needs of the neuroscience community. This facilitated a synergistic progress of both the research and infrastructure within EBRAINS. Moreover, the results reported herein also rely on the neural simulation tool NEST from EBRAINS’ Brain Simulation and simulation workflow Service Category (SC3) and dedicated High Performance Computing (HPC) resources provided by EBRAINS’ Interactive workflows on HPC or NMC Service Category (SC6).

In combination with the involvement with co-development of the NRP, the work shown in this Deliverable makes a significant contribution to one of the aims of Work-package 3, namely, to develop a large-scale functional cognitive architecture. The modularity of this architecture lends itself to convenient testing of various other visuomotor models in an embodied perspective. Moreover, it allows not only testing different implementations of the existing functions but also expanding on other functionalities. Additionally, the architecture is publicly available and hence accessible to all researchers interested in embodied cognitive architectures in general and scene understanding in particular.

2. Main achievements

2.1 Architecture

The architecture engages in a free viewing paradigm, making saccades based on the saliency of regions in the visual scene, identifying fixated objects and extracting their spatial relationship (Figure 2).

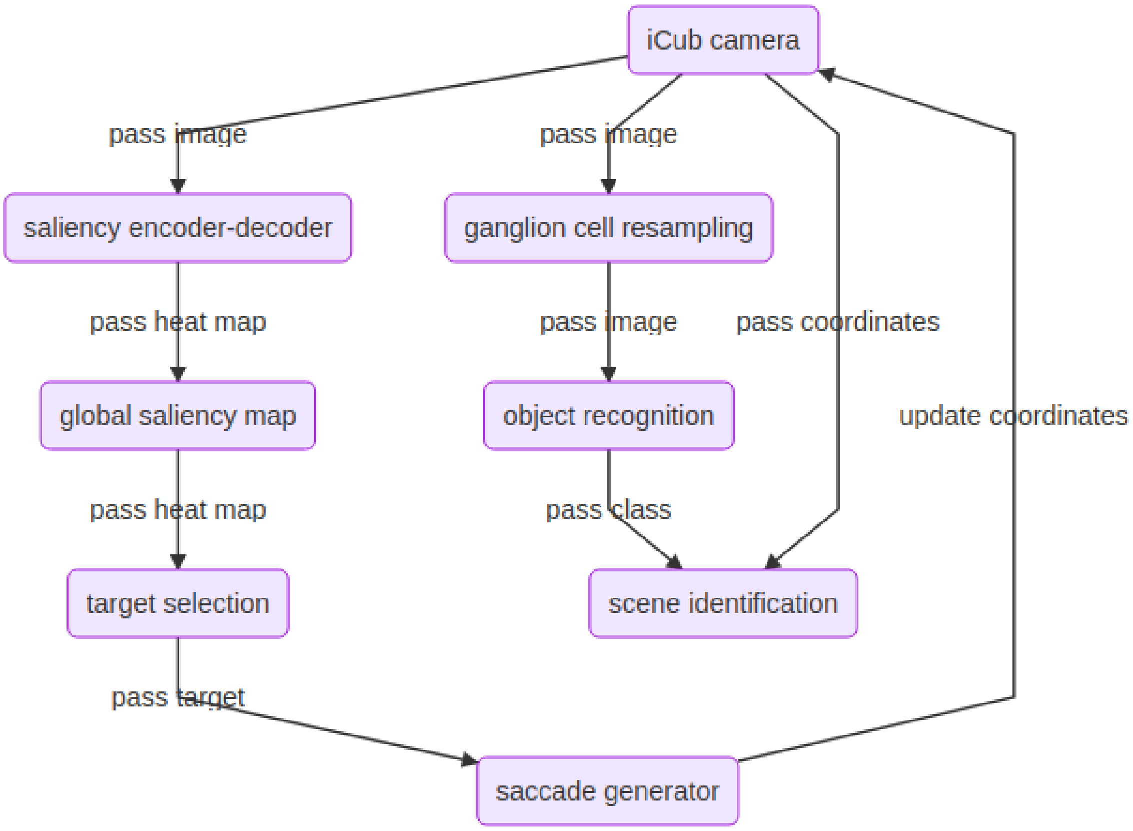

Figure 2: Closed-loop modular architecture for saccades for scene understanding

2.2 Models

Several models were developed for this Deliverable Details of each are listed below.

2.2.1 Ganglion image sampling

To take the physical constraints imposed by the eye into account, an embodied architecture of scene understanding needs to start with a retinal sampling of the visual scene. This can be achieved by resampling images captured by a robot’s camera according to the retinal ganglion cell distribution (da Costa et al., 2021). Specifically, the resampling operation converts between visual field coordinates and ganglion cell coordinates. Since ganglion cell placement is irregular foremost for eccentricity and less so for polar angles (Watson, 2014), resampling essentially is the effect of converting visual field radii to ganglion cell radii while keeping angles constant. Such a conversion allows one to move a pixel in the input image (camera) to its corresponding location in the output image. The resampled retinal image resembles a wheelbarrow distortion of the camera image in which the periphery is compressed while the fovea is stretched (Figure 3).

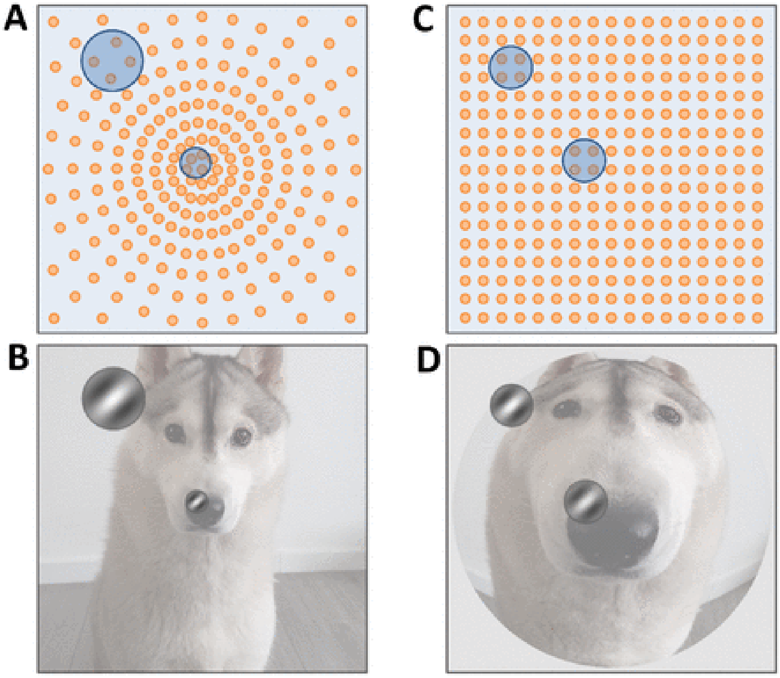

Figure 3: Schematic representation of retinal ganglion cell sampling. A, non-uniform distribution of ganglion cells. Spacing between cells increases as a function of eccentricity. B, differences in V1 receptive field sizes and spatial frequency tuning. (da Costa et al., 2021)

2.2.2 Saliency computation

To predict saliency from complex images, we leveraged high-level features extracted from natural scenes via a pre-trained VGG16 architecture and augmented them with contextual information derived from large receptive field sizes (Kroner et al., 2020). To that end, an Atrous Spatial Pyramid Pooling module (Chen et al., 2017) was introduced to the task of saliency prediction, which samples semantic information at multiple spatial scales and hence allows a more holistic assessment of an image’s content. Additionally, the influence of global scene context on fixation patterns (Torralba et al., 2006) was captured by pooling the activation of high-level features across the whole image. Together, this rich feature space formed the foundation for saliency prediction based on a convolutional decoder that restores the resolution of the visual input. The final encoder-decoder structure was trained on the SALICON and MIT1003 datasets to learn a mapping from raw pixel values to saliency maps that generalise to unseen examples. This network was then deployed in the NRP and received visual input from the iCub’s left eye camera. Individual saliency maps were integrated into a global representation of saliency beyond the current field of view, assuming a head-centred reference frame (Figure 4).

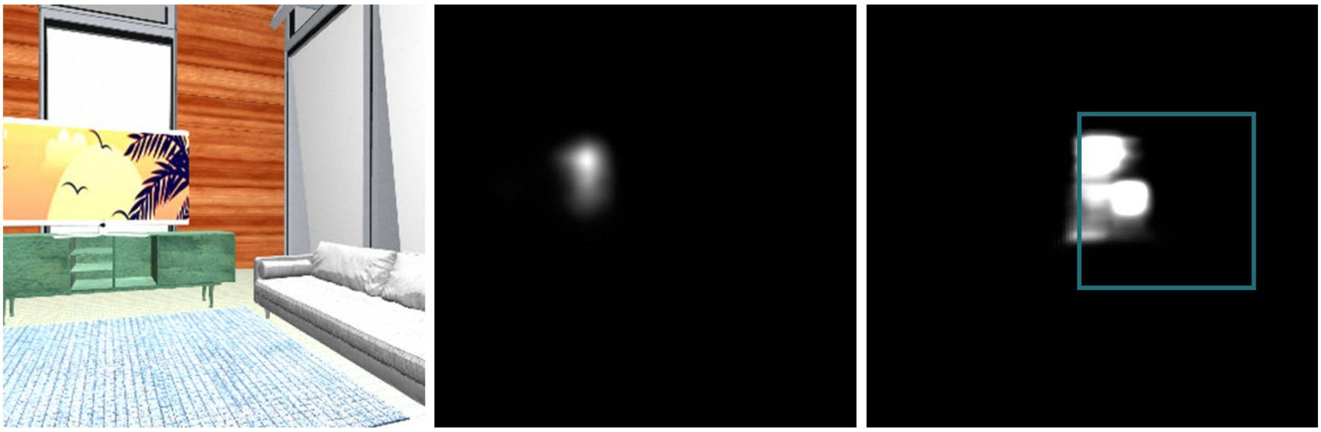

Figure 4: (left) A snapshot from the left eye of the iCub. (middle) The saliency map predicted from the current image. (right) The integration of local saliency into a global representation that stretches beyond the current field of view (blue square).

2.2.3 Target Selection

The target selection model presented here is a mean-field stochastic dynamical system, situated in the frontal eye-fields as they are involved in guiding saccades (Schall, 2004). The model generates a sequence of saccadic targets as seen in free-viewing conditions. We adopt the Generalised Lotka-Volterra Model (GLVM; Afraimovich et al., 2004) that can produce sequential dynamics to model neuronal activity. The GLVM equations were used to model a two-dimensional sheet of neuronal populations that process the visual salience map, computed by the saliency prediction model described above. On this sheet, inhibition strength between populations decay with increasing distance, i.e., close-by neighbours inhibit each other strongly. Each neuronal population in the two-dimensional network receives input corresponding to one location in the monocular visual world. This sheet of neuronal populations exhibits lateral inhibition and self-inhibition, which is crucial for the sequential dynamics required for this work (Afraimovich et al., 2004). Furthermore, the input dynamics to the populations incorporate firing rate-dependent inhibition of input. This is rooted in the phenomena of corollary discharge, an internal copy of the eye movement commands (Crapse & Sommer, 2008). It allows keeping track of fixations to enable inhibition of return to previously-visited locations (List & Robertson, 2007). The model is optimised (using SC6 HPC facilities) to satisfy certain behavioural constraints such as 3-5 saccades per second in a free-viewing paradigm (Walker, 2012) and that the scene is explored in a manner that is consistent with human free-viewing experiments (Wilming et al., 2017).

The model is described using the following equations:

$$\frac{dI_i}{dt} = -\frac{I_i}{\tau} + \mu\left(1 - \frac{\nu_i}{\nu^*}\right) I_{ext} + I_{noise}$$

$$\frac{d\nu_i}{dt} = \nu_i \left[I_i - \sum_{i=1}^{N} W_{i,j}\, \nu_j \right]$$

where I_i = input to the ith population in the sheet

ν_i = firing rate of the ith population in the sheet

τ = time constant of the input variable

μ = mean input to the population

ν^* = reversal firing rate; firing rate beyond which the input becomes inhibitory

I_{ext} = saliency distribution input

I_{noise} = noisy input; to introduce stochasticity

$W_{i,j}$ = connection strength from the jth population to the ith population

2.2.4 Saccade generator

The model of the saccade generator in the reticular formation presented here is a spiking neural network implementation of the Gancarz and Grossberg model (Gancarz & Grossberg, 1998). It consists of two identical components, one for horizontal and one for vertical saccades. Each component is further subdivided into two identical smaller circuits controlling an extra-ocular muscle each and thus being responsible for moving the eye in one direction. Each of these smaller circuits consists of five populations: long-lead burst neurons (LLBN), short-lead burst neurons (SLBN), inhibitory burst neurons (IBN), tonic neurons (TN), and omnipause neurons (OPN). The OPN population is shared by the two smaller circuits for either horizontal or vertical saccades. The distinction of these populations is motivated by different neuron types found in the reticular formation (Figure 5).

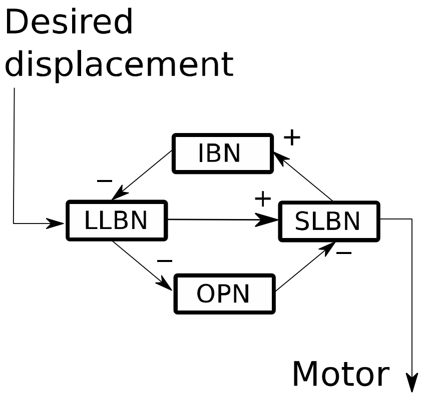

Figure 5: The saccade generator network (adapted from Gancarz and Grossberg, 1998) showing the various types of neurons interacting to generate a motor command. The figure here only represents the left side of the generator.

The LLBN and SLBN each consist of two interconnected populations of excitatory and inhibitory multi-timescale adaptive threshold model neurons (Kobayashi, 2009). This neuron model can generate burst neural firing. The OPN comprise two recurrently connected populations of excitatory and inhibitory leaky integrate-and-fire neurons. Finally, the IBN consist of one recurrently connected population of inhibitory multi-timescale adaptive threshold neurons, again showing bursting behaviour. All neurons receive independent Gaussian noise inputs. The noise is temporally uncorrelated (white) except for time discretisation at the simulation step size. Starting from this as a building block, the saccade generator, for e.g. horizontal saccades, is constructed by using two of these circuits — one for leftward and the other for rightward saccades — sharing an OPN population. The complete saccade generator then consists of two of the latter networks, one for horizontal and the other for vertical saccades. The full network is able to generate monocular, ballistic, eye movements in two dimensions. To that end, desired eye displacements obtained from the target selection model are converted into input into the four LLBN populations.

2.2.5 Scene identification

To recognise a scene, multiple snapshots are stacked over time in one sequence. Each snapshot is distorted using Ganglion-cell based distortion(da Costa et al., 2021). For each snapshot in a sequence, features are extracted using a pre-trained ConvNet. Then, an LSTM network (Hochreiter & Schmidhuber, 1997) is trained on the sequences of features to predict scene labels. For feature extraction, the InceptionV3 (Szegedy et al., 2016) network pre-trained on ImageNet is used. The features are extracted from an average pooling layer which is the last layer in the network before the classification layer. For each snapshot, a feature vector of length 2048, is generated.

The sequences of features are trained to predict labels using an LSTM network. The network consists of 1 LSTM layer with 1000 hidden nodes and 4 fully connected layers with 100, 50, 10 and 4 nodes, respectively. The last fully connected layer is a softmax classification layer (Figure 6).

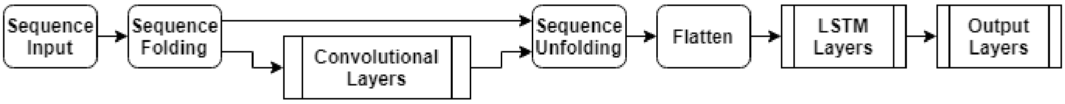

Figure 6: The scene identification architecture.

To make label predictions in an end-to-end manner, the pre-trained ConvNet is assembled with the LSTM network. Since the ConvNet only accepts individual snapshots and not sequences, a sequence folding layer is used. The sequence folding layer converts a sequence of images to a batch of images. After the features are extracted, the sequences (of features) are restored using a sequence unfolding layer and a flattening layer.

2.3 Embodiment on the Neurorobotics Platform

Since the iCub's camera image is fed through the different modules in sequential order, we decided to integrate all modules in one Transfer Function (TF) within the NRP. TFs are python scripts that interact with the simulation environment through receiving data from sensors and sending data to control robot actuators. The execution of TFs is synchronised with the world simulation, which is orchestrated by the NRP's Closed Loop Engine (CLE). This means that TFs are executed with a user-defined frequency, and during execution the world simulation is paused. In our case, we chose the TF timestep to be 5ms (200 hz) to match the simulation time of the Saccade Generator module.

For this experiment, the iCub is spawned in different rooms, and the TF is responsible for generating saccades that would help identify the correct room type. To realise this, different room layouts were defined within the NRP's Frontend. We created four different room types: bedroom, kitchen, living room, and office. And for each room type, we defined three different layouts. The layouts are yaml configuration files that specify a position and orientation for each object. If all objects under one layout definition are spawned in the NRP, it would yield a complete room in the Frontend.
Additionally, to create variability in the collected data, we gathered multiple models for each object type to be spawned in a room. Each time an object inside a room is spawned, a model is selected randomly from the model repertoire. This ensures that even if one room/layout pair is selected multiple times during data collection, the room would look different each time (Figure 7). Each layout also specifies multiple poses for the iCub, from which the perspective would allow for room identification.

To run the experiment, we used the Virtual Coach (VC), an NRP component that allows us to script and run batch NRP simulations in Python. With the VC, we launch and start the experiment and then select a random room/layout pair and spawn the objects in the corresponding positions. The spawning is done with the help of *rospy*, a Robot Operating System (ROS) Python package. After the room is ready, we parameterise the TF with the label, i.e. the room type, and add the TF to the simulation. Once the TF is added, it will start taking a snapshot every 5ms, feeding it through the Saliency, Target Selection and Saccade Generator modules, and move the eyes accordingly. After each movement, the new image, the current horizontal and vertical eye positions, and the label (room type) are saved to disk. This process is repeated for different room/layout pairs to create a large dataset. The number of iterations, as well as the number of images saved from each iteration, are user-defined parameters and can be adjusted in the VC script.

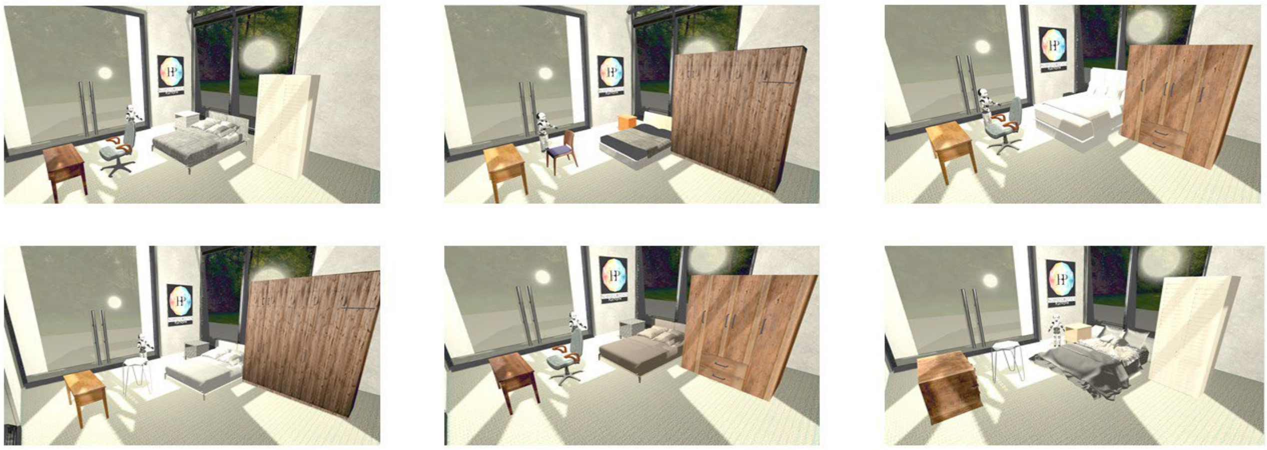

Figure 7: Six different versions of one bedroom layout within the Neurorobotics Platform.

3. Looking Forward

The closed-loop architecture described here provides an integrated neurocomputational model of saccades for scene understanding, including cortical and brain stem structures. The work conducted highlights a novel, holistic, approach to cognitive computational neuroscience characterised by the development of large-scale, embodied, neurocomputational models that are implemented in a modular fashion to allow for biological realism in conjunction with functional performance. Due to limited computational resources, modelling in neuroscience generally needs to trade-off between biological realism and functional performance. On the one hand, a focus on biological realism often implies focusing on a single or a few isolated neural structures performing simplified functions. On the other hand, a focus on the performance of ecologically valid functions often implies ignoring biological detail. However, individual neural structures and the functions they subserve can only be truly understood in the context of other neural structures and neurocomputational processes with which they interact and the input and output representations which constrain their computations. Similarly, reduced biological realism can have important functional implications as they constrain the fundamental neurocomputational operations that might be carried out by a neural structure. By developing large-scale architectures composed of modules which can individually be implemented at any degree of biological realism, this approach allows modelling any (rather than all) neural structure to a high degree of biological realism and embedding it in the larger architecture implemented with a lower degree of biological realism to ensure continued functional performance of the full system.

The present architecture provides a proof of concept and a first prototype upon which the development of architectures implementing high-level cognitive tasks can be built. This prototype will enable researchers within WP3 to study the performance and acquisition of perceptual, cognitive and motor tasks within the same unifying framework, to integrate detailed models of other structures such as the cerebellum or superior colliculus, as well as to systematically compare different models of the same neural structures (e.g., a different implementations of target selection in the frontal eye fields) within the same context.

References

Afraimovich, V. S., Zhigulin, V. P., & Rabinovich, M. I. (2004). On the origin of reproducible sequential activity in neural circuits. *Chaos: An Interdisciplinary Journal of Nonlinear Science*, *14*(4), 1123-1129. https://doi.org/10.1063/1.1819625

Chen, L.-C., Papandreou, G., Schroff, F., & Adam, H. (2017). Rethinking Atrous Convolution for Semantic Image Segmentation. *ArXiv:1706.05587 [Cs]*. http://arxiv.org/abs/1706.05587

Crapse, T. B., & Sommer, M. A. (2008). Corollary discharge across the animal kingdom. *Nature Reviews Neuroscience*, *9*(8), 587–600. https://doi.org/10.1038/nrn2457

da Costa, D., Goebel, R., & Senden, M. (2021). *ConvNets Develop Organizational Principles of the Visual Cortex when using Ganglion Cell-Based Sampling* [Preprint]. Neuroscience. https://doi.org/10.1101/2021.11.02.466130 - **P3081**

Gancarz, G., & Grossberg, S. (1998). A neural model of the saccade generator in the reticular formation. *Neural Networks*, *11*(7–8), 1159–1174. https://doi.org/10.1016/S0893-6080(98)00096-3

Hochreiter, S., & Schmidhuber, J. (1997). Long Short-Term Memory. *Neural Computation*, *9*(8), 1735–1780. https://doi.org/10.1162/neco.1997.9.8.1735

Kobayashi, R. (2009). Made-to-order spiking neuron model equipped with a multi-timescale adaptive threshold. *Frontiers in Computational Neuroscience*, *3*. https://doi.org/10.3389/neuro.10.009.2009

Kroner, A., Senden, M., Driessens, K., & Goebel, R. (2020). Contextual encoder–decoder network for visual saliency prediction. *Neural Networks*, *129*, 261–270. https://doi.org/10.1016/j.neunet.2020.05.004 - **P1761**

List, A., & Robertson, L. C. (2007). Inhibition of return and object-based attentional selection. *Journal of Experimental Psychology: Human Perception and Performance*, *33*(6), 1322–1334. https://doi.org/10.1037/0096-1523.33.6.1322

Schall, J. D. (2004). On the role of frontal eye field in guiding attention and saccades. *Vision Research*, *44*(12), 1453-1467. https://doi.org/10.1016/j.visres.2003.10.025

Szegedy, C., Vanhoucke, V., Ioffe, S., Shlens, J., & Wojna, Z. (2016). Rethinking the Inception Architecture for Computer Vision. *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2818–2826. https://doi.org/10.1109/CVPR.2016.308

Torralba, A., Oliva, A., Castelhano, M. S., & Henderson, J. M. (2006). Contextual guidance of eye movements and attention in real-world scenes: The role of global features in object search. *Psychological Review*, *113*(4), 766–786. https://doi.org/10.1037/0033-295X.113.4.766

Walker, J. (2012). Human saccadic eye movements. *Scholarpedia*, *7*(7), 5095. https://doi.org/10.4249/scholarpedia.5095

Watson, A. B. (2014). A formula for human retinal ganglion cell receptive field density as a function of visual field location. *Journal of Vision*, *14*(7), 15. https://doi.org/10.1167/14.7.15

Wilming, N., Onat, S., Ossandón, J. P., Açık, A., Kietzmann, T. C., Kaspar, K., Gameiro, R. R., Vormberg, A., & König, P. (2017). An extensive dataset of eye movements during viewing of complex images. *Scientific Data*, *4*(1), 160126. https://doi.org/10.1038/sdata.2016.126