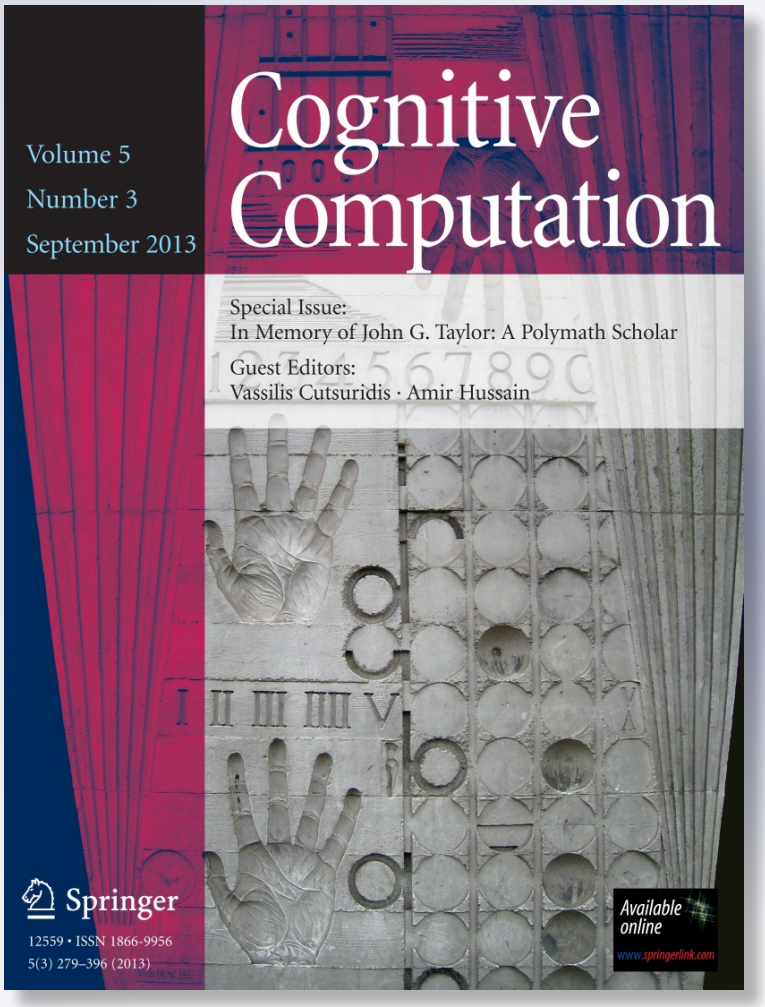
A Cognitive Control Architecture for the Perception–Action Cycle in Robots and Agents

Vassilis Cutsuridis & John G. Taylor

Cognitive Computation

ISSN 1866-9956
Volume 5
Number 3

Cogn Comput (2013) 5:383-395
DOI 10.1007/s12559-013-9218-z



Springer

Cogn Comput (2013) 5:383–395
DOI 10.1007/s12559-013-9218-z

A Cognitive Control Architecture for the Perception–Action Cycle in Robots and Agents

Vassilis Cutsuridis · John G. Taylor

Received: 11 September 2012 / Accepted: 2 April 2013 / Published online: 11 April 2013

Abstract We show aspects of brain processing on how visual perception, recognition, attention, cognitive control, value attribution, decision-making, affordances and action can be melded together in a coherent manner in a cognitive control architecture of the perception–action cycle for visually guided reaching and grasping of objects by a robot or an agent. The work is based on the notion that separate visuomotor channels are activated in parallel by specific visual inputs and are continuously modulated by attention and reward, which control a robot's/agent's action repertoire. The suggested visual apparatus allows the robot/agent to recognize both the object's shape and location, extract affordances and formulate motor plans for reaching and grasping. A focus-of-attention signal plays an instrumental role in selecting the correct object in its corresponding location as well as selects the most appropriate arm reaching and hand grasping configuration from a list of other configurations based on the success of previous experiences. The cognitive control architecture consists of a number of neurocomputational mechanisms heavily supported by experimental brain evidence: spatial saliency, object selectivity, invariance to object transformations, focus of attention, resonance, motor priming, spatial-to-joint direction transformation and volitional scaling of movement.

Keywords Saliency · Attention · Adaptive resonance theory · Decision-making · Value · Reaching · Grasping · Affordances · Perception–action cycle

V. Cutsuridis
Division of Engineering, Kings College London, Strand, London WC2R 2LS, UK

V. Cutsuridis (✉)
Foundation for Research and Technology—Hellas (FORTH), Nikolaou Plastira 100, 70013 Heraklion, Crete, Greece
e-mail: vcutsuridis@gmail.com; vcutsuridis@imbb.forth.gr

J. G. Taylor
Department of Mathematics, Kings College London, Strand, London WC2R 2LS, UK

Introduction

Industrial robotics has been successful in carrying out various types of operations such as assembling, dismantling and packaging in different areas of the mechanical, car, aerospace, automotive, electronics and other industries. Despite their successes almost all industrial robots still lack any form of human intelligence. Their capabilities are limited because their intelligence is passed onto them by a human programmer or engineer usually in the form of a list of step-by-step instructions executed in a continuous loop.

The majority of machine vision and object recognition systems today apply some form of mechanistic or deterministic template matching, edge detection or colour scanning approach for identifying and distinguishing different objects in a field of view. Objects are required to have a colour that provides a strong contrast with a background colour, in order to detect edges reliably. Detailed information about the location of objects in space is required in order to achieve the necessary reliability, accuracy of manipulation and effectiveness of an assembly process. Fine disturbances in the workspace of a robot can easily lead to failures. Their performance is slow and poor at identification, recognition, learning and adapting to noisy images and errors, compared to the human brain. They tend to deal only with one and only one task at hand

as they are constructed as single-purpose systems. It is very difficult to apply a current system to an even small variation in the same task.

Cognitive robotics, a relatively new branch of robotics, is attempting to address these limitations. It aims to endow the robot with intelligent behaviour by providing it with a processing architecture that will allow it to learn and reason about how to behave in response to complex goals in a complex world. The starting point for the development of such intelligent systems is human and animal cognition. Some of the cognitive capabilities these systems are empowered with include visual perception, attention, anticipation, planning, complex motor coordination and reasoning about the external world and perhaps even about their own mental states. Such systems are able to learn how to perform tasks automatically and adapt to unforeseen operating conditions or errors in a robust and predictable manner, without the need of human guidance, instructions or programming. Ultimately, these systems are embodied into a robot/software agent able to act in the real/virtual world.

In this paper we are introducing a cognitive control architecture of the perception–action cycle for visually guided reaching and grasping of objects. The objects themselves are not known a priori to the system, but their knowledge is assumed to be built by the system through interaction and experimentation with them. The architecture is multi-modular consisting of object recognition, object localization, attention, cognitive control, affordance extraction, value attribution, decision-making, motor planning and motor execution modules. The components of the architecture are novel as well as based on previously published works [3, 4, 13, 17, 19, 20, 26, 37, 62, 67] and follow very closely what we currently know of the human and animal brain. The model is tested against a hypothetical scenario where multiple objects are situated in the environment, and the robot/agent must recognize them, localize them, attend to each one of them and reach and grasp them according to an externally dictated sequence of motor actions.

Brain Pathways of the Perception–Action Cycle

The perception–action cycle has been defined by the eminent neuroscientist Joaquin Fuster as "the circular flow of information from the environment to sensory structures, to motor structures, back again to the environment, to sensory structures, and so on, during the processing of goal-directed behaviour" [27]. Between Fuster's sensory and motor structures, a number of areas exist, which are heavily interconnected, and each serves a particular function. An exhausting description of all these brain areas involved is impossible due to the lack of space and experimental evidence. In this section we will briefly describe only the human and animal brain areas we think are critical to the proper functioning of our perception–action cycle architecture.

The initial stages of visual processing involve the retina, lateral geniculate nucleus (LGN) and the primary visual cortex (V1). In these areas, individual neurons can discriminate small changes in visual orientations, spatial frequencies and colour [53]. V1 then transmits information to two pathways: the dorsal stream (the "where" pathway) and the ventral stream (the "what" pathway) [68]. The dorsal stream begins with V1 and goes through visual area V2, then to MT (visual area V5) and from there to the posterior parietal cortex (PPC). The dorsal stream, also known as the "where pathway", is associated with motion, representation of object locations and control of the eyes and arms, especially when visual information is used to guide eye movements or reaching [53]. The ventral stream begins with V1, to V2, then to visual area V4 and finally to the inferior temporal cortex (IT) [68]. The ventral stream, also known as the "what pathway", represents form recognition and object representation [45].

Another cortical area that receives input from the visual areas is the anterior intraparietal area (AIP). Experimental work has shown in anterior intraparietal area (AIP) the presence of a large number of neurons that are active in association with grasping and manipulation movements (motor neurons), presentation of visual stimuli (visual neurons) and both hand actions and object representations (visuomotor neurons) [49, 50, 59, 64]. AIP neurons receive direct input from the IT object-shape cells. Area AIP neurons transform the visual information of a given 3D object into multiple descriptions, thus providing the premotor areas with several grasping plans [28].

PPC and IT are known to project heavily to the prefrontal cortex (PFC), which in turn project back to these areas, thus linking it to perception, memory and action [48]. The prefrontal cortex is the anterior part of the frontal lobes of the brain, lying in front of the motor and premotor areas [53]. It has been implicated in planning complex cognitive behaviour, personality expression, decision-making and moderating social behaviour [2, 42, 46, 51]. It is considered to orchestrate thoughts and actions in accordance with internal goals [2, 42, 46, 51].

The premotor cortex (PM) is an area of motor cortex lying within the frontal lobe of the brain just anterior to the primary motor cortex [53]. It is involved in the planning of movements [53]. It is subdivided into two sections: the dorsal (upper) premotor cortex (PMd) and the ventral (down) premotor cortex (PMv). PMd plays a role in guiding reaching, whereas PMv in guiding grasping [53]. PMv contains the mirror neurons [56]. Mirror neurons are

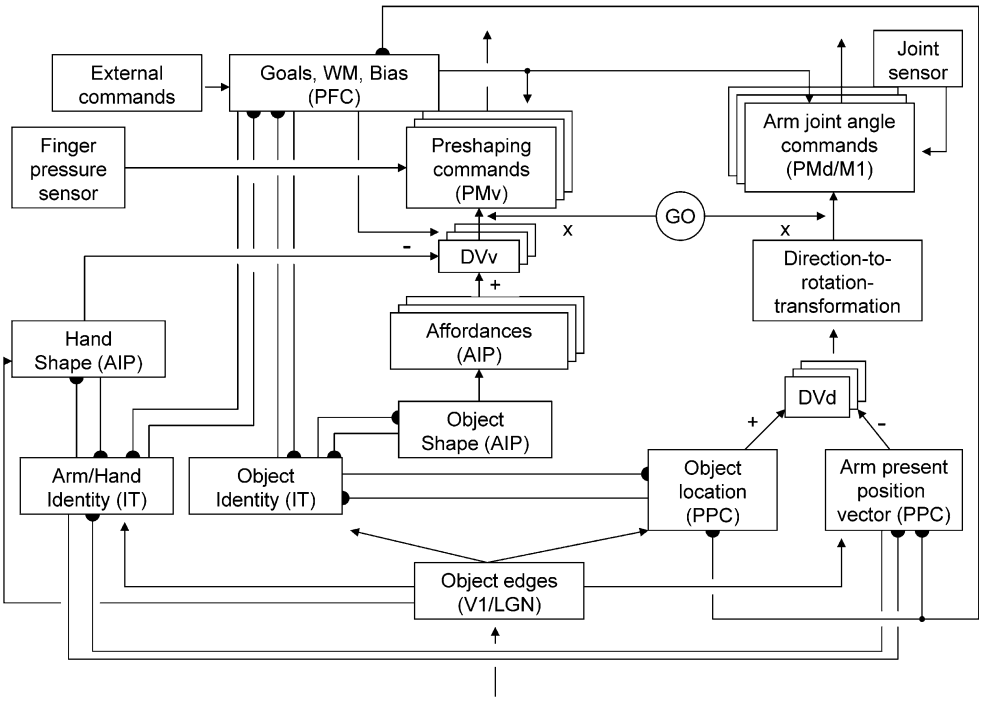


Fig. 1 Graphical representation of the complete cognitive control architecture of the perception–action cycle for visual-guided reaching and grasping

both sensory and motor as they become activated when an animal grasps an object as well as when the animal observes another animal grasps for an object [56]. Mirror neurons are proposed to be a basis for understanding the actions of others by internally imitating the actions using one's own motor control circuits [56]. PMv and AIP are strongly connected with each other.

The primary motor cortex (M1) is a brain region located in the posterior portion of the frontal lobe [53]. It is heavily connected with other motor areas including the premotor cortex and posterior parietal cortex, as well as several subcortical brain regions (e.g. basal ganglia structures), to plan and execute movements [53]. It is the area where the final motor command is formed before it is sent to the upper and lower extremities for execution [53].

Cortical areas are under constant modulatory control by neuromodulators such as dopamine (DA) and acetylcholine. Dopamine has been implicated in signalling value attribution and reward prediction errors used to select actions that will maximize the future acquisition of reward [61] as well as the progressive movement deterioration of patients suffering from Parkinson's disease [12, 14–16, 18]. Dopamine is produced by dopaminergic neurons located in the substantia nigra pars compacta (SNc) and the ventral tegmental area (VTA). Experimental evidence has shown that the SNc DA neurons are activated by neurons in the upper layers of the superior colliculus (an area responsible for the formation of the final motor command sent to the eyes), which in turn is activated directly by the retinal visual input [9, 11, 24, 47, 54]. The SNc DA neurons broadcast neuromodulatory signals to neurons in prefrontal cortex (PFC), premotor (PM) and primary motor (M1), posterior parietal (PPC) and temporal cortices (e.g. IT) [18, 70].

The Architecture

A graphical representation of our proposed architecture of the perception–action cycle is given in Fig. 1. The architecture proposes that acting upon objects in the environment like in the case of visually guided reaching and grasping involves two separate visuomotor channels, one for reaching and another one for grasping, which are activated in parallel by specific visual inputs, and each channel controls specific parts of limb (arm and hand, respectively). An input image is processed in a bottom-up fashion, providing input to feature detectors, which in turn lead to the formation of visual maps (the *object identity* map (ventral stream) and the *object location* (*spatial saliency)* map (dorsal stream)). Bidirectional crosstalk between object and spatial maps ensures the object corresponds to the appropriate spatial location in the environment. The visual maps then activate the *cognitive control* map (goals, motivations, task constraints), which in turn feeds back to amplify the neural representations in the visual maps, which are relevant to the current context, and to suppress the irrelevant ones. Adaptive resonance between "goals map" and "object map", and "goals map" and "spatial map" is achieved via a measure of degree of similarity, which depends on the amount of modulation (value attribution) the maps receive from the "value

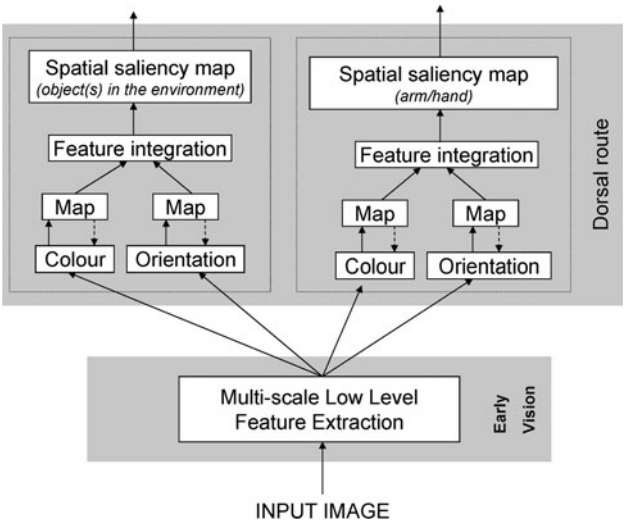


Fig. 2 Schematic of the information flow in the dorsal visual hierarchy for object localization (spatial saliency)

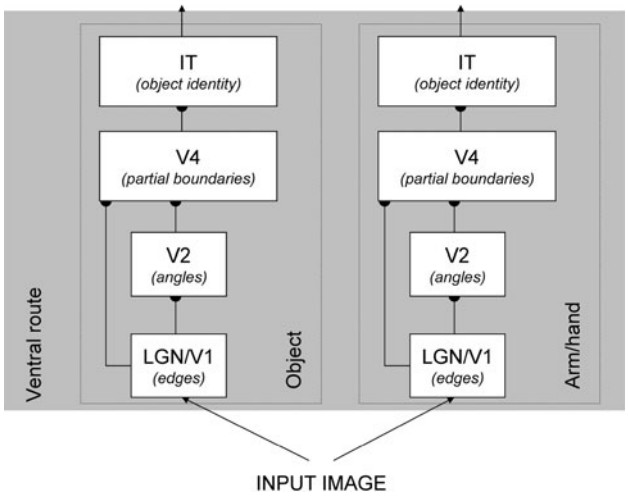


Fig. 3 Schematic of the information flow in the ventral visual hierarchy for 3D object identity recognition. LGN: lateral geniculate nucleus, V1 primary visual cortex, IT inferotemporal cortex

attribution" module. A winner-take-all competition between resonated neural representations ensures the object and spatial representations that reached resonance first will continue for further processing first, followed by the second fastest and so on. Once an object and a spatial representation is selected, then a library of action plans is selected, one for reaching and the other one for grasping. Once again, the cognitive control maps will select the action plan most relevant to the current context and suppress the irrelevant ones. The selected reaching and grasping motor plans will be gated by a volitional movement signal (GO) and form the final motor commands, which will be sent to the motor execution centres for execution. Feedback attention selects the most appropriate to the current context and externally received instruction motor command to move the arm/hand towards the target. Visual and proprioceptive feedback will update the current arm position and fingers configuration towards the desired ones.

In order for the architecture to achieve such complicated processes, a number of components are required. The topology, intra- and interconnectivity and proposed functionality are heavily supported by experimental brain evidence. We describe these components in the following sections.

Object Localization (Spatial Saliency)

The input image processing up to the formation of a spatial saliency map in the dorsal stream is similar to that described in Cutsuridis's [13] study, and it is depicted in Fig. 2. Its functionality is to decompose an input image through several pre-attentive multi-scale feature detection mechanisms into a set of distinct channels each tuned to a specific stimulus feature such as colour and orientation. Such decomposition is performed at a number of different spatial scales to allow for small and large objects to be represented in separate subdivisions of these channels. Different spatial scales are created using Gaussian pyramids [5], which consist of progressively low-pass filtering and subsampling of the input image. Pyramids with eight spatial scales are generated providing horizontal and vertical image reduction factors ranging from 1:1 (level 1; original image) to 1:256 (level 8) in consecutive powers of 2. A 3×3 Gaussian filter is applied to each level of the pyramid before the decimation operation yields the next level.

Local orientation information is obtained from the input image using oriented Gabor pyramids $O(\sigma, \theta)$ where σ takes values from [0, 8] and θ from $\{0°, 45°, 90°, 135°\}$; σ represents the scale and θ the preferred orientation. The Gabor filter [32] which is the product of a cosine function and a 2D Gaussian envelope, approximated the receptive field response of orientation-selective neurons in primary visual cortex [23, 52].

Colour information is obtained from the input image using the CIELab model [63]. In the CIELab model an RGB image is converted into the Lab colour space by ignoring variations in brightness. The Lab space consists of a luminosity layer "L", a chromaticity layer "a" indicating where colour falls along the red–green axis and a chromaticity layer "b" indicating where the colour falls along the blue–yellow axis. Since the colour information exists in the "ab" space, any objects in the image are pixels with "a" and "b" values. A KMEANS algorithm is then used to classify the objects into three categories (red, green and blue) using the Euclidean distance metric.

Each feature is computed in a centre-surround operation. Centre-surround operations are implemented as differences between a fine and a coarse scale for a given feature. Neurons in the feature maps in the dorsal stream then encode the spatial contrast in each of those feature channels. Neurons in each feature map spatially compete for salience, through long-range connections that extend far beyond the spatial range of the classical receptive field of each neuron. After competition, the feature maps are combined into a saliency map, which topographically encodes for saliency irrespective of the feature channel in which stimuli appeared salient. $N + 1$ spatial salient maps are computed: 1 for the arm/hand representation and N for the N potential objects in the environment to reach and grasp.

Object Recognition

Object recognition in our architecture is mediated by the ventral stream [68]. Our visual object recognition architecture is a hierarchical feedforward series of competitive networks using convergence from stage to stage (see Fig. 3). Convergence from a small part of a region to the succeeding one exists in such a way that receptive field size of neurons increases by a factor of 2.5 [58]. These convergence zones overlap extensively with each other [58]. The connections between the hierarchical competitive networks (from V1 to V2 to V4 to IT) are fixed (non-plastic), in order to increase the speed of simulations. In the competitive network at the top of the hierarchy (IT), an associative learning rule (Hebbian) associates the 2D object representations in order to construct a unified 3D object representation [58]. Each competitive network in our architecture is composed of simple (S) and complex (C) cells [53]. The first stage of processing is the simple units (S1) in V1. Each S1 cell is tuned in a Gaussian-like way to a bar of a certain orientation among a few possible ones. Each of the complex units (C1) in V1 receives the outputs of a group of S1 cells in V1 at slightly different positions, but with the same preferred orientation. The operation is performed via a nonlinear softmax operation, where the activity of a pooling unit corresponds to the activity of the strongest input, pooled over a set of synaptic inputs [69]. This increases invariance to local changes in position and scale while maintaining feature specificity.

At the next simple cell layer (S2) in V2, the S2 units pool the activities of several complex units (C1) in V1 with different selectivities according to a Gaussian tuning function, thus yielding selectivity to more complex patterns such as combinations of oriented lines forming angles [69]. Simple units (S3 and S4) in higher visual areas (V4 and IT) combine more and more complex features (partial boundaries, etc.) with a Gaussian tuning function, while the complex units (C2 and C3) pool their outputs through a max function providing increasing invariance to position and scale [69].

Value Attribution

Value in our architecture is broadcasted by the value attribution module as a modulation parameter (w) that selectively tunes the goals (PFC), spatial (PPC), object (IT) and motor programme (PMd/PMv) salient representations by increasing their signal-to-noise ratio and ensures their between resonance (see *decision-making* module for details). The value attribution module is activated by the model's low-level visual processing centres (not shown).

Cognitive Control

Cognitive control is the "ability to consciously manipulate thoughts and behaviours using attention to deal with conflicting goals and demands" [51]. In our architecture cognitive control neurons, which get activated by an external command as well as from the spatial map representations (PPC) and the object map representations (IT), set the order by which the goals, plans and actions are to be performed. It also receives input. Its role is to: (1) send a *focus-of-attention* signal to the spatial and object maps from which it receives input by amplifying the relevant context neuronal representations, while at the same time inhibiting those of distracters, and (2) participate in the adaptive resonance process (see next section) of the selectively tuned via the value attribution (see previous section) spatial, object and motor plan representations in their corresponding networks (PPC, IT, PMd and PMv) (see Fig. 4).

Decision-making

The decision to which object to reach and grasp next is determined by the coordinated actions of the *focus of attention*, *value* and *object and spatial maps* in the model (see Fig. 4). More specifically, bottom-up and top-down mechanisms represented by the complex and intricate feedforward, feedback and horizontal circuits of PFC, PPC and IT are making decisions. Adaptive reciprocal connections between (1) the "goals" map (PFC) and the "spatial" map (PPC), (2) the "goals" map (PFC) and the "object" map (IT) and (3) the "spatial" map (PPC) and the "object" map (IT) operate exactly as the comparison and recognition fields of an ART (adaptive resonance theory) system [7].

In its most basic form, an ART system consists of two interconnected fields of neurons: the comparison field and the recognition field. The comparison field responds to input features, whereas the recognition field responds to categories of the comparison field activity patterns.

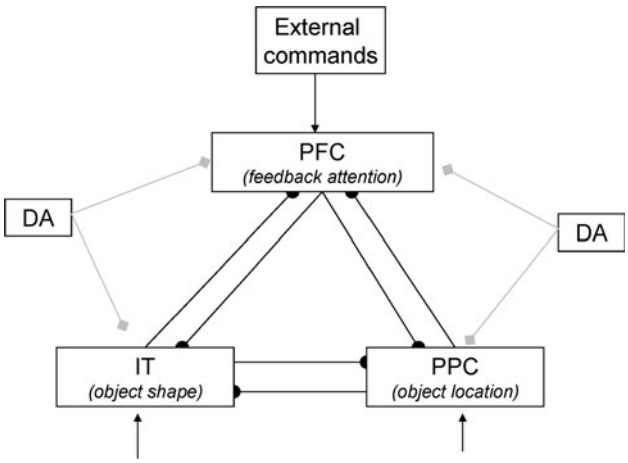


Fig. 4 Schematic of the information flow of the model's decision-making processes. Bidirectional adaptive connections among the "goals" map (PFC) and the "object" map (IT), and the "goals" map (PFC) and the "spatial" map (PPC) work like an ART network. Value attribution signals from the value attribution module (SNc) to the "goals" (PFC), "object" IT and "spatial" (PPC) maps act like ART's vigilance parameter, selectively tuning the goals, object and spatial map representations

Bidirectional connections between the two fields are adaptive (modifiable). Neurons in the recognition field compete with each other in a recurrent on-centre off-surround fashion. Inhibition from the recognition field to the comparison field shuts off most of the comparison field activity, if the input mismatches the active category's response. If the match is close, enough of the comparison field nodes excited by both the input and the active category node overcome the non-specific inhibition of the recognition field. If on the other hand mismatch occurs, the recognition field inhibition shuts off the active category node as long as the current input is present. Matching occurs when sufficient correspondence between comparison and recognition field patterns is greater than a parameter value called vigilance.

In the model, the ART's vigilance parameter is represented by the broadcasted value attribution signals (see previous section). High and intermediate value levels ensure the formation of fine and coarse categories, respectively, whereas low values (low signal-to-ratio signals) ensure that non-relevant representations and plans perish. The reciprocal connections between goals map (PFC), spatial map (PPC) and object map (IT) allow for the amplification of the spatial and object representations pertinent to the given context and the suppression of the irrelevant ones, whereas the reciprocal connections between spatial map (PPC) and object map (IT) ensure for their groupings.

Affordances

In our architecture affordances are the potential finger preshaping configurations that will best allow the object to be grasped successfully. In our architecture affordances are represented by the AIP neuronal map. AIP neurons receive direct input from the IT object-shape cells. Area AIP neurons transform the visual information of a given 3D object into multiple descriptions, thus providing the premotor areas with several grasping plans [28]. The premotor areas (specifically PMv) then select the most appropriate based on the contextual information it receives from cognitive control module (PFC).

Movement Planning and Execution

Reaching is achieved by a self-organizing neural network model capable of solving the inverse kinematics problem by spatially localizing the learning task within the vicinity of each robot configuration reached during learning [4, 38, 39]. Our reaching neural network (see Fig. 5) achieves its competence by transforming visual information about hand target position (PPC object location) and actual hand spatial position (PPC hand location) into a body-centred spatial representation of the direction in 3D space that the hand must move to locate the hand near the object to be grasped. The spatial direction vector (DVd) is adaptively transformed into a motor direction vector through the direction-to-rotation transform, which represents the synergistic joint rotations of shoulder and elbow that move the wrist in the desired spatial direction from the present arm configuration. A volitional movement signal (GO) gates the direction-to-rotation transform motor direction vector and generates a list of potential final motor commands. A focus-of-attention signal is sent from the "goals" map (PFC) neurons to select the most appropriate final motor reaching command to be expressed and suppress all others.

For the grasping process the Vector-Integration-To-Endpoint (VITE) model [3, 67] is used to model the finger preshaping neural channel involved in prehension (see Fig. 6). VITE gradually integrates the difference vector (DV) between the desired target hand postures (affordances) arising from the AIP cell activities and retrieved from the external library of finger configuration gestures and the actual finger configuration described by the arm/hand AIP activities. The rate of integration (i.e. the movement velocity) is controlled by the product of the DV vector and a volitional movement gating signal (GO) arising from the basal ganglia structures. As before, the "goals" map cells then sent a focus-of-attention signal to the most appropriate final grasping motor command to be expressed, while suppressing all others.

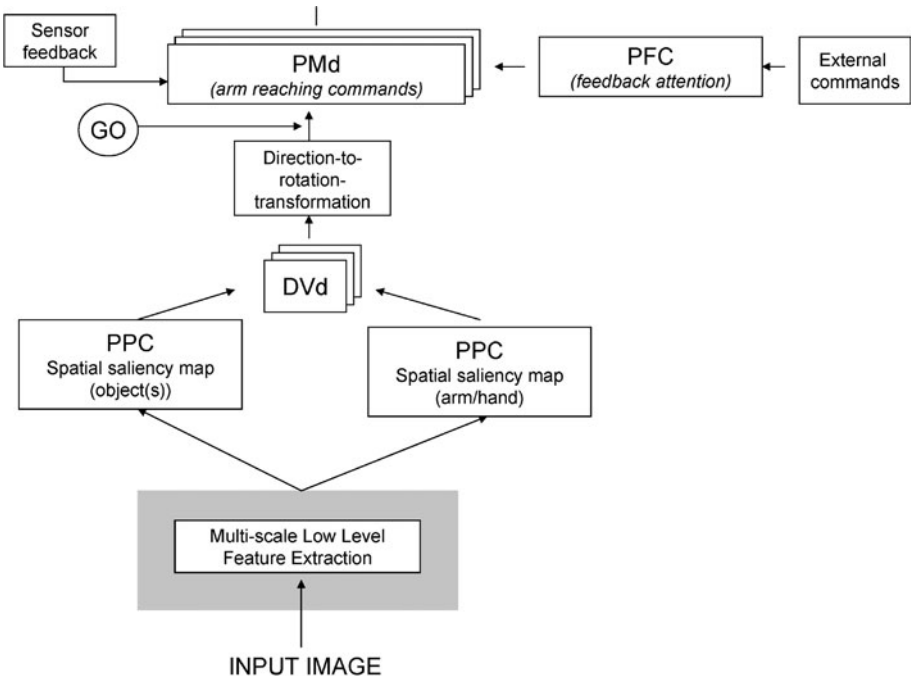


Fig. 5 Schematic of the information flow of the arm/hand transport component

Bringing Everything Together

In this section we will describe how all components of the architecture described above work together when multiple objects are situated in the environment and the robot/agent must recognize them, localize them, attend to each one of them and reach and grasp them according to an externally dictated sequence of motor actions. In order for the robot/agent to tackle each scenario successfully, additional constraints need to be imposed on the robot/agent and task: The robot/agent is situated in front of a table, and it is unable to move its torso (i.e. it is immobile). The robot's head is facing the table, and it is not allowed to move left or right, up or down. Its eyes are also not moving (passive vision). The robot/agent is allowed to move only one of its arms/hands, but not both. The robot's arm/hand is also considered an object that the robot is attempting to move towards the stationary object(s) in the environment in order to reach for it/them and grasp it/them. On the table an object or multiple objects are situated at different locations, which are not known a priori by the robot/agent. The objects may have different shapes, sizes and colours. The

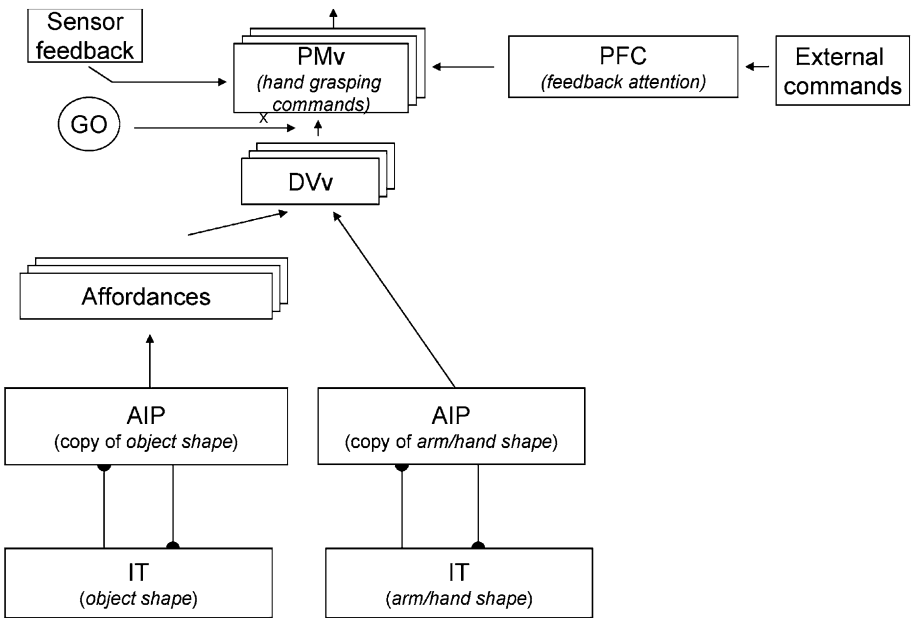


Fig. 6 Schematic of the information flow of the object grasping component

behavioural task for the robot/agent is to execute a sequence of reach-to-grasp actions, dictated by external commands, towards an object in the environment.

Objects in the Environment

When an image frame of the video stream is presented to the robot's eyes, two parallel and equally fast-processing modes of actions are initiated. In the first mode of action (object localization (spatial saliency) and object recognition processing; see Fig. 7a), pre-attentive multi-scale feature detection and extraction mechanisms sensitive to different features (colour and orientation) operating in parallel at the level of the retina, LGN and V1 start to work in order to recognize the identity and location of the objects (including arm/hand) in the environment. From the level of V1 and on, the features are separated into two streams: the dorsal for object location (spatial saliency) processing (see Fig. 7a left) and the ventral for object identity processing (see Fig. 7a right). As we mentioned earlier, along the ventral stream information is processed in a feedforward way through a series of competitive neural networks, where the receptive field size of neurons in each succeeding state increases by a factor of 2.5. At the first level (V1) of processing, lines of different orientations (0, 45, 90, 135) are extracted via Gabor filters, which are then combined in V2 to form angles. At the level of V4, lines from V1 and angles from V2 are combined to form partial boundaries. At the last level (IT), partial boundaries are added to form 2D object-shape representations. At IT the different 2D object representations are associated via a Hebbian learning rule to form 3D object representations (see Fig. 7b, [58]. Complex cells at each level of processing (V1 to V4) ensure that their neuronal representations are invariant to local changes in size, position and scale, while they maintain feature specificity.

Along the dorsal stream and at the first level (V1) of processing, local orientation information and colour information are obtained from the input image via Gabor filters and the CIELab model, respectively (see Sect. 3.1 for details). Each feature is computed in a centre-surround operation, where the difference between a fine and a coarse scale for a given feature is computed. Neurons in the feature maps then encode the spatial contrast in each of those feature channels. Neurons in each feature map spatially compete for salience, through long-range connections that extend far beyond the spatial range of the classical receptive field of each neuron. After competition, the feature maps are combined into a spatial saliency map, which topographically encodes for spatial saliency irrespective of the feature channel in which stimuli appeared salient.

Bidirectional associative connections between the spatial saliency (PPC) and object identity (IT) maps ensure that the identified object is associated with its corresponding location in the environment.

In the second mode of action, the low-level visual processing centres activate the value attribution module, which in turn broadcast the value attribution (DA) signals to the goals (PFC), spatial (PPC) and object (IT) map representations and selectively tune the responses of different neuronal populations in these areas according to the previous similar acquired experiences. ART search cycles [7] will begin where top-down attentional signals from PFC will send/receive top-down/bottom-up feedback/feedforward signals to/from the spatial (PPC) and object (IT) maps (see Fig. 7c). The value attribution signals acting as a similarity measure between the top-down goals map (PFC) and bottom-up spatial saliency (PPC) and object identity (IT) map representations will selectively tune the neuronal responses at the PFC, PPC and IT levels. A winner-take-all mechanism in these fields will select those spatial and object representations that reached resonance with the PFC ones. These selected PPC and IT representations will continue to the next stages of processing.

In the next step two parallel modes of action will be performed: one for reaching (see Fig. 7e) and the other for grasping (see Fig. 7d). For reaching, the two spatial (PPC) neuronal representations (N for N objects and 1 for the arm/hand) that successfully reached resonance will be subtracted from each other, constructing this way N difference vectors (DVd) in spatial coordinates. The difference vectors will be then transformed into a motor direction vector through the direction-to-rotation transform that will move the arm/hand in the desired spatial direction from the present arm configuration. A volitional movement signal (GO) will gate the direction-to-rotation transform motor direction vectors and generate N final motor commands. Feedback attention from PFC will select the most appropriate to the current context final motor command to drive the arm/hand towards the desired object location. Proprioceptive information from the robot/agent joints will fine-tune the movement of arm/hand towards the final target location.

For grasping, the IT 3D object-shape neuronal representation that reached resonance will be copied to the AIP 3D neuronal map. The copied AIP representation will extract from an external library a list of possible target finger preshaping configurations (affordances) that had in the past led to a successful grasp of the target object. Each of these affordance representations will be subtracted from the AIP actual finger configuration (hand aperture), thus forming a series of difference vectors (DVv). Each difference vector will be gated by a volitional GO signal, thus generating a series of final grasping commands. Top-down attentional signals from the PFC will bias the most

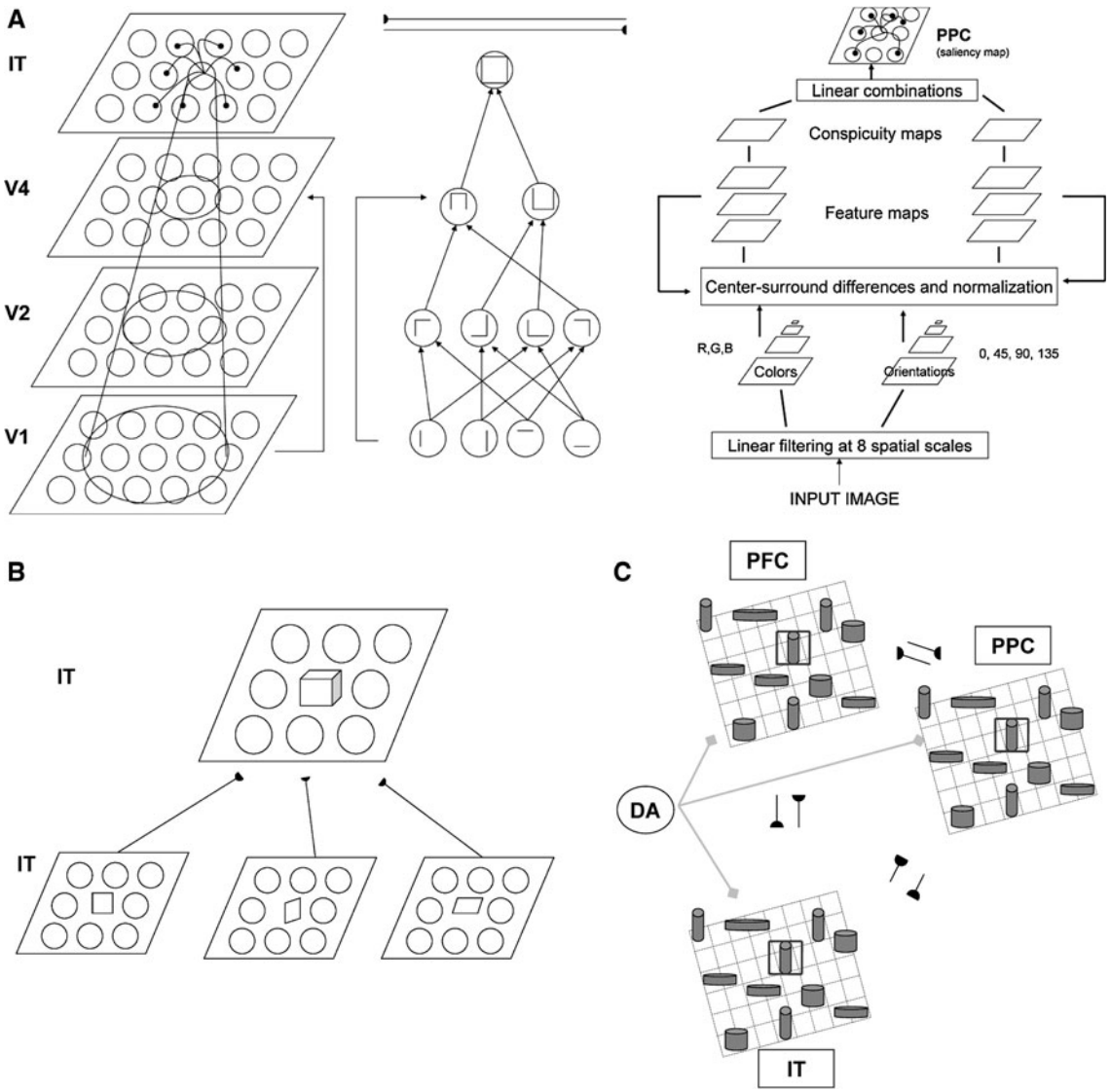


Fig. 7 **a** Visual hierarchy as it occurs in the ventral (*left*) and dorsal (*right*) streams in the architecture. (Ventral stream) V1 primary visual cortex tuned to lines of different orientation, V2 tuned to angles, V4 tuned to partial boundaries and IT inferior temporal cortex tuned to object recognition. (Dorsal stream) PPC posterior parietal cortex tuned to object location. Bidirectional adaptive connections between IT and PPC ensure that the recognized object corresponds to its recognized spatial location. **b** Weighted combination of 2D object identity representations form 3D object representation in the level of IT. **c** Value attribution signals to spatial (PPC), goals (PFC) and object identity (IT) maps. Different neuronal populations in these maps receive different levels of value modulation. High and intermediate value modulation result in "sharp tuned" neuronal responses, whereas low-value modulation results in "broadly tuned" neuronal responses. Neuronal responses are depicted by *grey*-coloured towers in each brain. The height of each tower represents the neuronal amplitude activation, whereas the width of each tower represents the degree of tuning. *Dark grey square* surrounding the response of a neuronal population represents the winner neuronal population from the ART search cycle, where resonance between the top-down (e.g. PFC) and bottom-up (e.g. PPC) populations according to some values of the vigilance (value modulation signal) has been achieved. **d** Grasping process. The winner IT object and arm/hand neuronal populations are copied to AIP. The AIP "Copy of object shape" population activates all potentially desired finger configurations (affordances) capable of grasping the object. The "actual finger" configuration is then subtracted from each "desired" one, thus generating *N* "difference vectors" (DVv). Each DVv is then multiplied by a volitional movement scaling signal (GO), thus generating N "hand grasping" commands. Feedback attention signal from PFC then selects the most appropriate hand grasping command according to the context dictated by the "external commands". The process continues till DVv is zero. **e** Reaching process. Spatial saliency maps are generated at the level of PPC for both the "object(s)" and the "arm/hand". The object spatial map is the "desired spatial map", whereas the arm/hand spatial map is the "actual spatial map". Both spatial maps are subtracted to form *N* difference vectors (DVd) in spatial coordinates. Each DVd vector undergoes a spatial-to-joint transformation, which is then scaled by a volitional signal (GO) and forms *N* "arm reaching motor commands". The most appropriate arm reaching motor command is then selected by the attention feedback PFC signal. The reaching process continues till DVd is zero

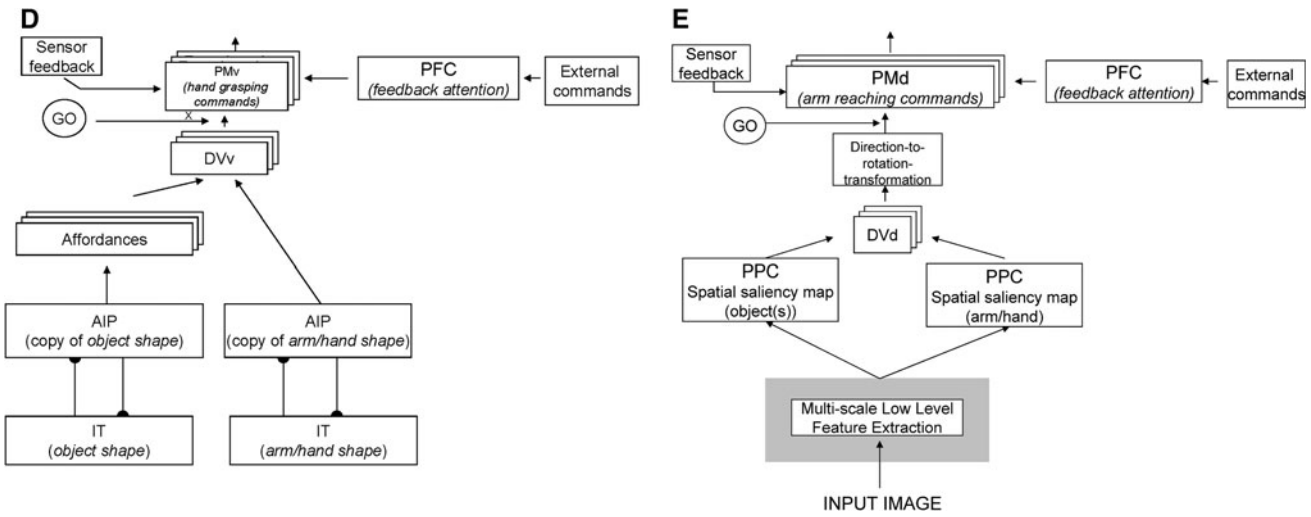


Fig. 7 continued

appropriate motor command, thus making it the most likely one to be executed.

When both DVd and DVv become zero, then the arm/hand will have successfully reached and grasped the object. The above-described process takes place for each image frame of the video stream.

Discussion

What Have We Learned from the Model?

The model presented in this paper is a cognitive architecture of vision-based reaching and grasping of objects located in the immediate environment of a robot/agent. The architecture proposes separate visuomotor channels, which are activated in parallel by specific visual inputs control the robot's/agent's action repertoire. The proposed robot's/agent's visual apparatus allows it to recognize both the object's shape and location, extract affordances and formulate motor plans for reaching and grasping. A focus-of-attention signal plays an instrumental role on selecting the correct object in its location as well as selects the most appropriate arm reaching and hand grasping configuration from a list of other configurations based on the success of previous experiences. Several classes of mechanisms have been detailed: *spatial saliency*, *object selectivity*, *invariance to object transformations*, *focus of attention*, *resonance, motor priming, spatial-to-joint direction transformation* and *volitional scaling of movement*.

Model's Mechanisms are Neuroscientifically Sound

The *spatial saliency* mechanism measured the level of conspicuity each object location had on a spatial map via image decomposition and normalization and neuronal competition [13]. Neural substrates of saliency maps have been found throughout the cortex including the posterior parietal cortex (spatial map representation in our model), frontal eye fields (motor programme representation in our model) and the prefrontal cortex (goal map representation in our model) as well as in subcortical areas such as the pulvinar and the superior colliculus [8, 10, 33, 43, 57, 66].

The *object selectivity* and *invariance to object transformations* mechanisms operated via deep-belief networks consisting of simple and complex neurons. These networks performed a series of nonlinear softmax operations ensuring in each stage of processing invariance to local changes in position and scale, while maintaining feature specificity [62]. The existence of simple and complex cells in V1 has been reported by Hubel and Wiesel [36]. Simple cells are selectively tuned via a Gaussian function to a particular stimulus feature (e.g. a particular orientation) [23]. Complex cells performing a softmax pooling operation (invariance to translation) have been reported in V1 [44] and V4 [29]. The object recognition component of the architecture has been able to duplicate quantitatively the generalization properties of IT neurons that remain highly selective for particular objects, while being invariant to some transformations [45].

The *focus-of-attention* mechanism included the more specific mechanisms of *amplification* of relevant information and the *suppression* of irrelevant ones throughout the visual and the motor programmes fields. Experimental [8, 55, 60] and computational [13, 22, 65] studies have confirmed the presence of such a signal in the brain.

The *resonance* mechanism worked as the matching process between the representations of the object, spatial and goals-in-context maps based on the *focus-of-attention* mechanism generated by the goals module and the *value*

attribution mechanism, which worked like a vigilant parameter of an ART network [7]. The representations that reached resonance first were the ones that were processed further first, followed by the second fasters and so on.

Motor priming operated in the form of a difference vector (DVd and DVv in the architecture) unable to cause any overt movement has been observed in the motor areas of monkeys [31]. A signal similar to the *volitional movement* signal (GO signal) has been shown in the data of Horak and Anderson [34, 35] where the activity of the internal segment of the globus pallidus was an in vivo analogue of the model's GO signal. Multiplication of the difference vector by the GO signal generated the motor command sent to lower motor centres for reaching and grasping. The *spatial-to-joint direction transformation* mechanism has been reported in the results of several neurophysiological experiments in monkey brains. A number of experimental investigators have reported cell activities in primary motor cortex, premotor cortex and supplementary cortex tuned to both external space and joint direction [1, 6, 30, 40, 41].

Future Work

Work is currently underway to implement the current model on the iCub robot as part of the EC-funded DARWIN project. The DARWIN project aims at developing robotic technology that will be able to assemble objects from their constituent parts. This assembly process will operate in three modes: (a) slave mode, where the necessary sequence of operations will be provided by a CAD-CAM system; (b) semi-autonomous, where the sequence will be provided either through demonstration by a teacher performing the same task or by describing it in a higher-level language, suitable for a human operator; and (c) fully autonomous mode, where an executive process will generate the necessary sequence by reasoning and mental simulation of consequences of actions on objects. The embodied architecture will be then tested against a number of assembly tasks with various levels of difficulty (e.g. insert an object into another and close the container object (i.e. a simplistic packaging scenario) or assemble a scaled model of the Eiffel tower or a scaled down version of a real industrial assembly process in a factory floor).

In addition, extensions of the model presented in this paper are currently under development. A particularly interesting extension of the model is how it may resolve conflicts when two actions receive the same amount of bias from PFC or have an equal probability of being expressed, because they have reached resonance at the same time. Cutsuridis et al. [21] have shown that such conflict resolution can occur via a simple competition between decision signals. Recent experimental evidence has shown that conflict resolution may also be resolved more centrally in anterior cingulate and prefrontal cortices [25]. Another extension of the model is a concept system allowing the formation of representations for the learning of spatial relations (up, down, inside, outside, etc.), the learning of causal relations among the features of concepts and allowing abstractions of objects in order to support the generalization performance of the system. Finally, another interesting extension is enhancement of motor capabilities supporting the coordination of a body with a pair of two arms that will be used to assemble the object in question.

References

1. Alexander GE, Crutcher MD. Neural representations of the target (goal) of visually guided arm movements in three motor areas of the monkey. J Neurophysiol. 1990;64(1):164–78.
2. Braver TS, Cohen JD. On the control of control: the role of dopamine in regulating prefrontal function and working memory. In: Monsell S, Driver J, editors. Control of cognitive processes: attention and performance XVIII. Cambridge: MIT Press; 2000. p. 713–38.
3. Bullock D, Grossberg S. Neural dynamics of planned arm movements: emergent invariants and speed-accuracy properties during trajectory formation. Psychol Rev. 1988;95:49–90.
4. Bullock D, Grossberg S, Guenther F. A self organizing neural model for motor equivalent reaching and tool use by a multijoint arm. J Cogn Neurosci. 1993;5(4):408–35.
5. Burt PJ, Adelson EH. The Laplacian pyramid as a compact image code. IEEE Trans Commun. 1983;31:532–40.
6. Camintini R, Johnson P, Urbano A. Making arm movements within different parts of space: dynamic aspects in the primate motor cortex. J Neurosci. 1990;10:2039–58.
7. Carpenter GA, Grossberg S. Adaptive resonance theory. In: Arbib MA, editor. The handbook of brain theory and neural networks. 2nd ed. Cambridge: MIT Press; 2003. p. 87–90.
8. Chelazzi L, Duncan J, Miller EK, Desimone R. Responses of neurons in the inferior temporal cortex during memory guided visual search. J Neurophysiol. 1998;80(6):2918–40.
9. Coizet V, Comoli E, Westby GW, Redgrave P. Phasic activation of Substantia Nigra and the ventral tegmental area by chemical stimulation of the superior colliculus: an electrophysiological investigation in the rat. Eur J Neurosci. 2003;17(1):28–40.
10. Colby CL, Goldberg ME. Space and attention in parietal cortex. Ann Rev Neurosci. 1999;22:319–49.
11. Comoli E, Coizet V, Boyes J, Bolam JP, Canteras NS, Quirk RH, Overton PG, Redgrave P. A direct projection from the Superior Colliculus to substantia Nigra for detecting salient visual events. Nat Neurosci. 2003;6(9):974–80.
12. Cutsuridis V. Does abnormal reciprocal inhibition lead to co-contraction of antagonist muscles? A modeling study. Int J Neural Syst. 2007;17(4):319–27.
13. Cutsuridis V. A cognitive model of saliency, overt attention and picture scanning. Cognit Comput. 2009;1:292–9.
14. Cutsuridis V. Neural network modeling of voluntary single joint movement organization. I. Normal conditions. In: Chaovalitwongse WA, Pardalos P, Xanthopoulos P, editors. Computational neuroscience. Berlin: Springer; 2010. p. 181–92.
15. Cutsuridis V. Neural network modeling of voluntary single joint movement organization. II. Parkinson's disease. In:

Chaovalitwongse WA, Pardalos P, Xanthopoulos P, editors. Computational neuroscience. Berlin: Springer; 2010. p. 193–212.
16. Cutsuridis V. Origins of a repetitive and co-contractive pattern of muscle activation in Parkinson's disease. Neural Networks. 2011;24(6):592–601.
17. Cutsuridis V. (2012). The perception-…-action cycle cognitive architecture and autonomy: a view from the brain. J Artif General Intell (in press).
18. Cutsuridis V, Perantonis S. A neural model of Parkinson's disease bradykinesia. Neural Netw. 2006;19(4):354–74.
19. Cutsuridis V, Heida T, Duch W, Doya K. Neurocomputational models of brain disorders. Neural Netw. 2011;24(6):513–4.
20. Cutsuridis V, Hussain A, Taylor JG. Perception-action cycle: Models, architectures and hardware. USA: Springer; 2011.
21. Cutsuridis V, Smyrnis N, Evdokimidis I, Perantonis S. A neural network model of decision making in an antisaccade task by the superior colliculus. Neural Networks. 2007;20(6): 690–704.
22. Desimone R, Duncan J. Neural mechanisms of selective visual attention. Ann Rev Neurosci. 1995;18:193–222.
23. DeValois RL, Albrecht DG, Thorell LG. Spatial-frequency selectivity of cells in macaque visual cortex. Vis Res. 1982; 22:545–59.
24. Dommett E, Coizet V, Blaha CD, Martindale J, Lefebre V, Walton N, Mayhew JE, Overton PG, Redgrave P. How visual stimuli activate dopaminergic neurons at short latency. Science. 2005;307(5714):1476–9.
25. Egner T, Hirsch J. Cognitive control mechanisms resolve conflict through cortical amplification of task relevant information. Nat Neurosci. 2005;8(12):1784–90.
26. Fagg AH, Arbib M. Modelling parietal-premotor interactions in a primate control of grasping. Neural Netw. 1998;11(7–8):1277–303.
27. Fuster JM. Upper processing stages of the perception-action cycle. TICS. 2004;8(4):143–5.
28. Gallese V, Fadiga L, Fogassi L, Luppino G, Murata A. A parietal-frontal circuit for hand grasping movements in the monkey: evidence from reversible inactivation experiments. In: Their P, Karnath HO, editors. Parietal lobe contributions to orientation in 3D space. Berlin: Springer; 1997. p. 255–70.
29. Gawne TJ, Martin JM. Responses of primate visual cortical V4 neurons to simultaneously presented stimuli. J Neurophys. 2002;88:1128–35.
30. Georgopoulos AP, Kalaska JF, Crutcher MD, Camintini R, Massey JT. The representation of movement direction in the motor cortex: single-cell and population. In: Edelman GM, Gall WE, Cowan WM, editors. Dynamic aspects of cortical function. New York: Wiley; 1984. p. 501–24.
31. Georgopoulos AP, Schwartz AB, Ketter RE. Neuronal population coding of movement direction. Science. 1986;233:1416–9.
32. Gonzalez RC, Woods RE. Digital image processing. New Jersey: Prentice Hall; 2002.
33. Gottlieb JP, Kusunoki M, Goldberg ME. The representation of visual salience in monkey parietal cortex. Nature. 1998; 391(6666):481–4.
34. Horak FB, Anderson ME. Influence of globus pallidus on arm movements in monkeys. I. Effects of kainic acid induced lesions. J Neurophysiol. 1984;52:290–304.
35. Horak FB, Anderson ME. Influence of globus pallidus on arm movements in monkeys. I. Effects of stimulations. J Neurophysiol. 1984;52:305–22.
36. Hubel DH, Wiesel TN. Receptive fields and functional architecture of monkey striate cortex. J Phys. 1968;195:215–43.
37. Itti L, Koch C. A saliency based search mechanism for overt and covert shifts of visual attention. Vis Res. 2000;40:1489–506.
38. Jordan MI. Motor learning and the degrees of freedom problem. In: Jeannerod M, editor. Attention and performance XIII: Motor representation and control. Hillsdale: Erlbaum; 1990. p. 796–836.
39. Jordan MI, Rumelhard DE. Forward models: supervised learning with a distal teacher. Cogn Sci. 1992;16:307–54.
40. Kalaska JF, Cohen DA, Hyde ML, Prud'homme M. A comparison of movement direction related versus load direction related activity in primate motor cortex using a two dimensional reaching task. J Neurosci. 1989;9(6):2080–102.
41. Ketter RE, Schwartz AB, Georgopoulos AP. Primate motor cortex and free arm movements to visual targets in three dimensional space. III. Positional gradients and population coding of movement direction from various movement origins. J Neurosci. 1988;8(8):2938–47.
42. Kravitz DJ, Saleem KS, Baker CI, Mishkin M. A new neural framework for visuospatial processing. Nat Rev Neurosci. 2011; 12:217–30.
43. Kusunoki M, Gottlieb J, Goldberg ME. The lateral intraparietal area as a salience map: the representation of abrupt onset, stimulus motion and task relevance. Vis Res. 2000;40:1459–68.
44. Lampl I, Ferster D, Poggio T, Riesenhuber M. Intracellular measurements of spatial integration and the MAX operation in complex cells of the cat primary visual cortex. J Neurophys. 2004;92:2704–13.
45. Logothetis NK, Pauls J, Poggio T. Shape representation in the inferior temporal cortex of monkeys. Curr Biol. 1995;5: 552–63.
46. MacDonald AWI, Cohen J, Stegner V, Carter CS. Dissociating the role of dorsolateral prefrontal and anterior cingulated cortex in cognitive control. Science. 2000;288(5472):1838–935.
47. McHaffie JG, Jiang H, May PJ, Coizet V, Overton PG, Stein BE, Redgrave P. A direct projection from superior colliculus to substantia Nigra Pars compacta in the cat. Neurosci. 2006;138(1): 221–34.
48. Miller E. The prefrontal cortex and cognitive control. Nat Rev Neurosci. 2000;1:59–65.
49. Murata A, Gallese V, Kaseda K, Sakata H. Parietal neurons related to memory guided hand manipulation. J Neurophys. 1996;75:2180–6.
50. Murata A, Gallese V, Luppino G, Kaseda K, Sakata H. Selectivity for the shape, size and orientation of objects for grasping in neurons of monkey parietal area AIP. J Neurophysiol. 2000;83: 339–65.
51. O'Reilly R, Braver T, Cohen J. A biologically based computational model of working memory. In: Miyake A, Shah P, editors. Models of working memory: mechanisms of active maintenance and executive control. Cambridge: Cambridge University Press; 1999.
52. Palmer S. Vision science: photons to phenomenology. USA: MIT Press; 1999.
53. Purves D, Augustine GJ, Fitzpatrick D, Hall WC, LaMantia AS, McNamara JO, White LE. Neuroscience. USA: Sinauer Associates Inc; 2004.
54. Redgrave P, Gurney K. The short latency dopamine signal: a role in discovering novel actions. Nat Neurosci. 2006;7:967–75.
55. Reynolds JH, Desimone R. The role of neural mechanisms of attention in solving the binding problem. Neuron. 1999;24(1): 19–29.
56. Rizzolatti G, Sinigaglia C. The functional role of the parieto-frontal mirror circuit: interpretations and misinterpretations. Nat Rev Neurosci. 2010;11(4):264–74.
57. Robinson DL, Petersen SE. The pulvinar and visual salience. TINS. 1992;15(4):127–32.
58. Rolls E. Memory, attention and decision making: a unifying computational neuroscience approach. Oxford: Oxford University Press; 2008.
59. Sakata H, Taira M, Murata A, Mine S. Neural mechanisms of visual guidance of hand action in the parietal cortex of the monkey. Cereb Cortex. 1995;5:429–38.

60. Schall JD, Hanes DP, Thompson KG, King DJ. Saccade target selection in frontal eye field of macaque. I Visual and premovement activation. J Neurosci. 1995;15:6905–18.
61. Schultz W. Predictive reward signal of dopamine neurons. J Neurophys. 1998; 80:1–27.
62. Serre T, Kouh M, Cadieu C, Knoblich U, Kreiman G, Poggio T. A theory of object recognition: computations and circuits in the feedforward path of the ventral stream in primate visual cortex. CBCL Memo 259, MIT (2005).
63. Sharma G. Digital color imaging handbook. New York: CRC Press; 2003.
64. Taira M, Mine S, Georgopoulos AP, Murata A, Sakata H. Parietal cortex neurons of the monkey related to the visual guidance of hand movement. Exp Brain Res. 1990;83:29–36.
65. Taylor JG, Hartley M, Taylor N, Panchev C, Kasderidis S. A hierarchical attention-based neural network architecture, based on human brain guidance, for perception, conceptualisation, action and reasoning. Image Vis Comput. 2009;27:1641–57.
66. Thompson KG, Bichot NP. A visual saliency map in the primate frontal eye field. Prog Brain Res. 2005;147:251–62.
67. Ulloa A, Bullock D. A neural network simulating human reach-grasp coordination by continuous updating of vector positioning commands. Neural Netw. 2003;16(8):1141–60.
68. Ungerleider LG, Haxby JV. 'What' and 'where' in the human brain. Curr Opin Neurobiol. 1994;4:157–65.
69. Wiersing H, Koerner E. Learning optimized features for hierarchical models of invariant object recognition. Neural Comput. 2003;15(7):1559–88.
70. Williams SM, Goldman-Rakic PS. Widespread origin of the primate mesofrontal dopamine system. Cereb Cortex. 1998;8: 321–45.