Computational model of enactive visuospatial mental imagery using saccadic perceptual actions

Jan Jug[0] & Tine Kolenik[0]
University of Ljubljana, Slovenia
André Ofner[0]
University of Vienna, Austria
Igor Farkaš
Comenius University in Bratislava, Slovakia

Abstract

From the onset of cognitive revolution, the concept of mental imagery has been given different, many times opposing, theoretical accounts. Mental imagery appears to be a ubiquitous, yet wholly individual, easy to explain experience on the one hand, being hard to deal with scientifically on the other hand. The focus of this research is on an enactive approach to visuospatial mental imagery, inspired by Sima's perceptual instantiation theory. We designed a hybrid computational model, composed of a forward model, an inverse model, both implemented as neural networks, and a memory/controller module, that grounds simple mental concepts, such as a triangle and a square, in perceptual actions, and is able to reimagine these objects by performing the necessary perceptual actions in a simulated humanoid robot. We tested the model on three tasks – salience-based object recognition, imagination-based object recognition and object imagination – and achieved very good results showing, as a proof of concept, that perceptual actions are a viable candidate for grounding the visuospatial mental concepts as well as the credible substrate of visuospatial mental imagery.

Keywords: enaction, mental imagery, visuospatial cognition, saccades, cognitive robotics

1. Introduction

Mental imagery (MI) is a phenomenon that has been given multiple (many times opposing) theoretical accounts from the start of the cognitive revolution, being tackled by such prominent figures as Pylyshyn (1973,

[0]These authors contributed equally to the work. They were supported by Erasmus+ scholarship during their mobility semester.

2002), Fodor (1975), Block (1981), Kosslyn (1980, 1994) and Barsalou (1999). The plethora of research on the topic is grounded in the fact of MI being an ubiquitous, yet wholly individual experience on the one hand, and easy to explain, yet hard to deal with scientifically on the other hand. A textbook definition (Eysenck, 2012) paints MI as the representation in a person's mind of the physical world outside of that person. It is characterized as a quasi-perceptual experience, as it occurs in the absence of what is perceived to be the appropriate stimuli from the outside. Aside from representing such a rich element in our mental lives, it is thought to be central to many cognitive abilities, such as memory (Paivio, 1986) and motivation (McMahon, 1973), but its foremost role is its involvement in visuospatial reasoning (Sima, 2014) and creative thought (Palmiero et al., 2016). The former is the focus of our own research.

There are many approaches to researching visuospatial MI, both theoretical and methodological. There are three prevailing theories: the pictorial theory (Kosslyn, 1994), the descriptive theory (Pylyshyn, 2002) and the enactive theory (Thomas, 1999). The pictorial theory claims that MI is the processing of the mental image in the visual buffer using processes of visual perception. This visual buffer is supposedly used in a parallel way during visual perception in order to create a mental representation of what is perceived. The descriptive theory claims that MI is the processing of amodal descriptions, which constitute the mental image. These descriptions are not a part of, or processed by, sensorimotor-related mechanisms. The enactive theory claims that MI emerges with the use of the same schemata that are used for perceiving the external world, e.g., certain schemas of eye movements. For instance, the well known Soar symbolic cognitive architecture, extended with a spatial visual system and a mental imagery module (Lathrop & Laird, 2009) has features of pictorial and descriptive theories, but not the enactive theory.

The enactive theory will be described more in-depth, as it serves as a paradigm for this research. Methodologically, analytic and synthetic approaches to science (Mirolli & Parisi, 2009) are both valid when researching MI (Sima, 2014). The analytic approach to science constitutes researching a phenomenon through observation and experiment. Cognitive psychology (Chambers & Reisberg, 1985), cognitive neuroscience (Bartolomeo & Chokron, 2002) and phenomenology (Thompson, 2007) have dealt with MI in this way. The synthetic approach to science tries to understand phenomena by making computer or robot models. The approach tries to apply principles, used and learned from successful implementations of computer models, to explain real phenomena. It sees models as possible explanations of reality. More specifically, one of the most common methods in modeling cognitive phenomena is the use of artificial neural networks (ANNs), which serve as a bridge between behavior and biology (O'Reilly & Munakata, 2000). ANNs were used in this research as well.

The paper is organized as follows. Section 1 provides of an overview of enactive approaches to mental imagery, including perceptual instantiation theory (Sima, 2012), that serves as the main conceptual source for our work. Section 2 presents the architecture of our model. Section 3 presents the simulations of the proposed model on three specified tasks. Section 4 describes the results of simulations. Section 5 provides the discussion of the model performance and the potential extensions. Section 6 summarizes the paper.

1.1. Enactive approaches to vision

The fundamental movement that spawned enactive sensorimotor approaches was the ecological cognition movement. One of the most important concepts from it is Neisser's (1976) schema, conceptualized to account for his idea of cognition, especially perception. According to Neisser, organisms don't just pick up information from the environment, they actively search for the information they need from the environment. Schemata serve to explain how organisms extract needed information. Organisms use participatory schemata to select information by constructing anticipations of information and waiting for the information to occur in the environment. Only then can information be acquired. Neisser's notion summarizes this: "We can see only what we know how to look for" (Neisser, 1976, p. 20). Therefore, there is a direct relation between perception and action. Schemata are a part of the perception-action cycle: schemata direct action to information, which is picked up by action and goe to schemata, modifying it in the process.

Neisser's account is somewhat consistent with the well-known ecological approach to visual perception (Gibson, 1986). It similarly focuses on researching how an active agent extracts information from the environment. Gibson also rejects the idea that sensory inputs are simply transformed into perceptions by some processes in the mind, and strongly advocates that perception can only be explained in terms of active observers, especially observers that move (or, more accurately, perform a motor action). Perception is therefore by definition not passive. The most relevant concept from Gibson's approach for the means of this research is the idea of affordances. Simply stated, an affordance is what environment affords or offers the agent. In more applicable terms, it is especially connected to categorization. By taking affordances seriously, categories can be defined by actions affording the perceptions of a specific category.

Arbib (1981) relies on Gibsonian ecological psychology and Neisser's concepts to offer his account on the phenomena, heavily shaped by cybernetics and control theory. He unambiguously characterizes perception "as potential action" (Ibid., p. 1459) through the concept of action-perception cycles, saying: "The subject's exploration of the visual world is directed by anticipatory schemas, which Neisser defines as plans for perceptual action as well as readiness for particular kinds of optical structure. The information picked up modifies the perceiver's anticipations of certain kinds of information that, thus modified, direct further exploration and prepare the perceiver for more information" (Ibid., p. 1458).

These approaches were most prominently followed by a more contemporary enactive, sensorimotor theory of perceptual consciousness (O'Regan's and Noë's, 2001). A similar idea emerges as before – that sensory stimulation depends on an active agent, on a perceiver in action. However, O'Regan and Noë attribute more power to action, as they don't believe that acting is only for retrieving sensory information – it equally contributes to perception itself as a whole, as experience.

Another aspect, not directly present in enactive visual perception accounts, yet clearly related, is the construction of our personal visual world and the role of saccades in this process. A saccade is a very fast movement of both eyes from one position to another. There are up to 5 saccades per second occurring in every individual (Hancock et al., 2012). This movement is not smooth, it is rather a jump, and it is done unconsciously. It is also consciously undetected due to its speed and top-down visual processing that constructs the world we see (Blackmore et al., 1995). The latter is necessary to build this conscious visual model of the world that we experience, otherwise we would experience the perceived visuals as constantly going in and out. We also do not take in the whole rectangular picture before us as experienced bottom-up – it is only due to saccades that go from position to position that we can construct this stable, whole image. This may also be a crucial difference between biological visual perception and computer vision. While biological vision constructs the experienced image one bit at a time through fast moving saccadic movements, computer vision takes in the picture in front of the camera as a whole (Figure 1).

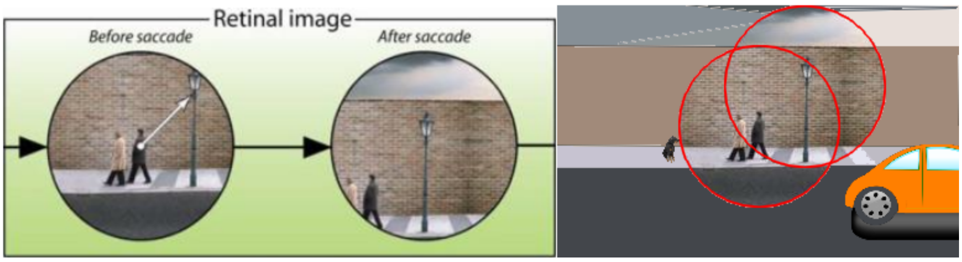


Figure 1: Left (Bays & Husain, 2008): The visual percept we take in in order to construct the experienced picture of the world. The left bit is one salient object (the man), the right bit contains another salient object (the lamp). Right: The approximal picture of the world we experience, constructed top-down from visual memory and other processes. To construct it, saccades are needed to other salient objects, like the dog and the car, therefore at least 3 saccades (man → lamp → dog → car). This also represents the picture that computer vision immediately perceives, without the need of biological construction (Szeliski, 2011).

These aspects of visual perception contribute to the understanding of the enactive approach to MI and its applicability in this research.

1.2. *Enactive approaches to mental imagery*

The first comprehensive account for the enactive approach to mental imagery was realized by Thomas (1999). It does not only encompass visual perception, but all perceptual modalities. The theory can be condensed into four principles: 1) mental representations do not exist as such, 2) perception is realized by actively interrogating the environment, 3) agents possess unique perceptual instruments for interrogating the environment for information and extracting it, 4) these perceptual instruments are guided by the agents' schemata. For illustration, consider looking at another person. The observer considers bottom-up information, which guides, and is in turn, guided by the top-down schemata. With perceptual instruments, the person perceives them as a whole (with saccades, among other things). Then the agent closes his eyes. Schemata for a person guide appropriate perceptual instruments (saccades, among other things) and try to recognize the person, but there is no person. This causes mental imagery. Sima (2014) builds upon this theory with his perceptual instantiation theory of visuospatial theory. Our work is essentially based on this approach.

1.3. *Perceptual instantiation theory*

Sima's perceptual instantiation theory (PIT) incorporates enactive approaches to visual perception (discussed previously) and the studies on eye movements. Along with aspects of these (especially relevant to this research is the notion that recognition is successful using top-down guided perceptual actions (PAs) to external stimuli; PAs will be discussed later on), the main assumption is that perceptual processes are "re-used" in MI. There is a number of mechanisms, connected with both visual perception and MI. The construction of the visual world is affected by bottom-up, external stimuli, which is realized in the agent as so-called perceptual information, but there is also top-down involvement, namely more conceptual information, coming from mental concepts. Mental concepts hold conceptual information, which may be qualitative, i.e. "red, small, square" and the necessary guidelines for enacting the right PAs (for the concepts in question; used in MI, but also in anticipation and prediction of the external world). The case of PAs is central to PIT. They are all those movements that enable the extraction of information from the environment (in case of visual perception, these are saccades and micro-saccades, lens adjustment, head movements, etc.). The main point of PAs is therefore retrieving the needed information from the environment, and different kinds of PAs can retrieve different kinds of information.

Another important aspect of Sima's theory is the visuospatial long-term memory (VS-LTM). It serves as a glue between mental concepts and PAs, as it maps one onto another and vice versa in order to produce the knowledge of how to look at the world and recognize entities in it. This constitutes a long-term memory, while a more general short-term memory serves as a keeper for current perception: identified mental concepts, perceptual information and the interpretation of the two merged together (what we see as a whole – e.g., when perceptual information is retrieved, it is compared to plausible mental concepts and the most consistent one is chosen for interpretation, which guides the PAs to retrieve even more information). Mental imagery supposedly builds on most of these concepts. For mental imagery, mental concepts are utilized and with the help of VS-LTM engage appropriate PAs. However, since there are no external stimuli from the environment and no perceptual information that guides the mental concepts (at least consciously), we do not get a picture of the real world, but rather a mental image, yet produced with a set of similar (mostly unconsciously driven) bodily movements as when perceiving (saccades, lens adjustment, etc.). After MI comes into place, perceptual information can be retrieved from it, and this, according to Sima, then leads to high-level cognitive processes, like reasoning.

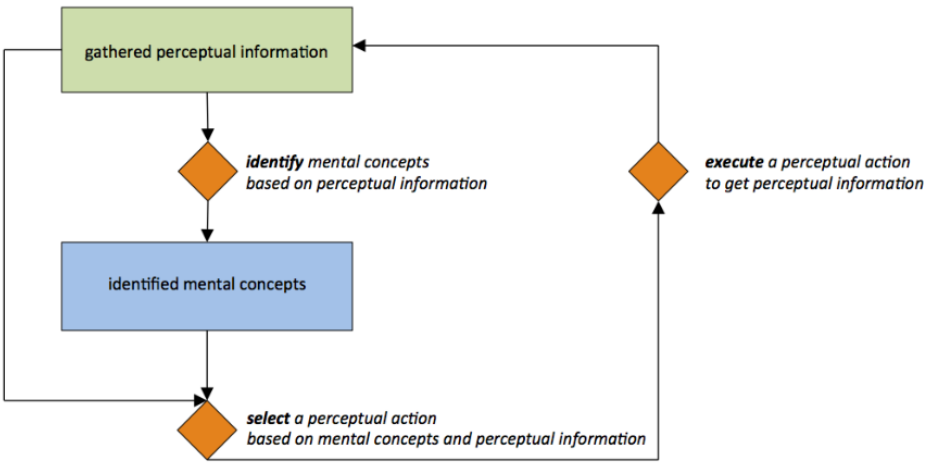


Figure 2: Visual perception and MI cycle: "1) the selection of a PA based on the identified mental concepts and available perceptual information; 2) the execution of the PA to retrieve further perceptual information; and 3) the identification of mental concepts based on the available perceptual information" (Sima, 2014, p. 70).

Last but not least, yet another extremely important aspect of PIT is that it has actually been formalized, which is a big departure from most previous, to a degree too vague and abstract discussions on MI. The basis of PIT is the formal description of the MI operands: a) perceptual information: low-level features that agents can perceive (edges, color, etc.), b) perceptual actions: basic actions of agents' visual system (saccades, lens adjustments, etc.), c) mental concepts: conceptual information, linking perceptual information and PA. These operands function in a cycle, shown in Figure 2. This cycle is further incorporated into a formal framework of PIT, as can be seen in

Figure 3. Sima also presents a computational model, but it is completely symbolic and does not come with implementation.

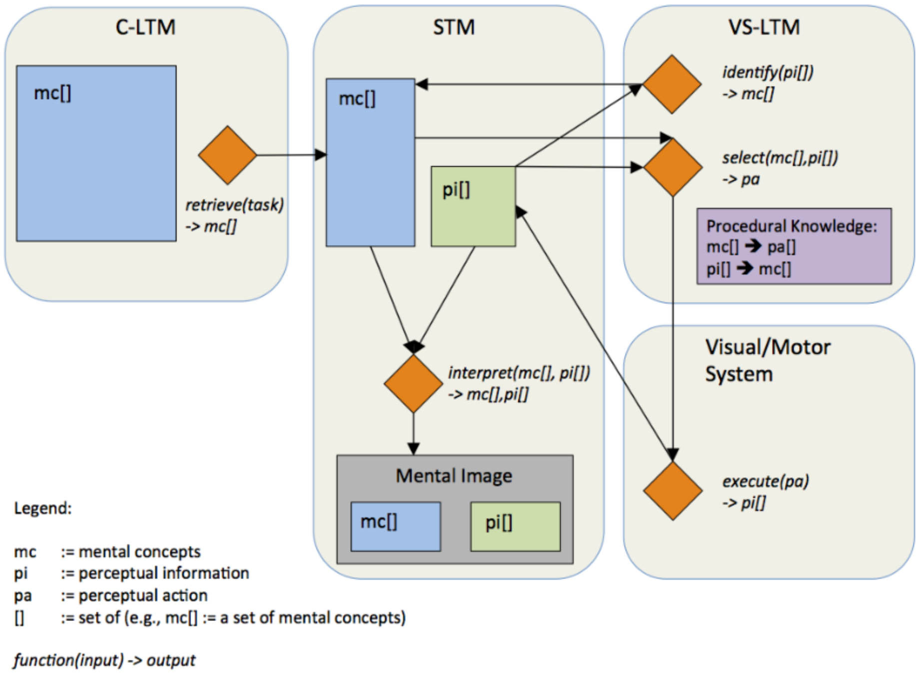


Figure 3: Mental imagination: "1) the retrieval of a set of mental concepts from C-LTM (long-term memory of conceptual information) which conceptually describe the scene; 2) these mental concepts are successively instantiated with perceptual information by the cyclic process of select-execute-identify; 3) an interpretation is drawn from all identified mental concepts with their instances of perceptual information; 4) this interpretation constitutes the mental image of the scene" (Sima, 2014, p. 71).

1.4. Our model

We take the main ideas of PIT, supplement them with our own and implement them in a biologically more relevant artificial neural network model. The most innovative contributions of our research include the novel work on robot vision with the inclusion of research on saccades and construction of the visual world (which called for improvisation in regards to limiting the usual visual field of robot vision), merged with enactive aspects on visual perception and MI (e.g., the meeting of bottom-up and top-down mechanisms, PAs, affordances in relation with mental concepts).

The ANNs are often used in controlling the iCub, one of the most accurate child-like robots, which has 53 degrees of freedom, movable eyes with cameras and numerous other sensors. The simulation of the iCub, used for the research, is built on Open Dynamics Engine, which provides a safe and ecological environment for testing. Our own testing for the iCub and its enactive visual and mental image characteristics is based on actual cognitive neuroscientific work to ensure ecological validity. This especially includes findings on salience in regards to movements of saccades – namely that when going from object A to the most salient object B, there is some kind of inhibition to avoid loops, i.e. going back to the most salient object from object B, which would be object A (Hooge & Frens, 2000), that saccades land towards center-of-mass position (Findlay, 1982) – but also the work on edges (of, for example, shapes), recognized by, e.g., shading (Humphrey et al., 1996). There is also evidence that eye movements during mental imagery are not epiphenomenal but assist the process of image generation. In other words, the eye scanpaths during visual imagery reenact those of perception of the same visual scene, therefore playing a functional role (Laeng & Teodorescu, 2002; Bourlon et al, 2011).

The role of perceptual actions (albeit called with different names) has also already been proved to be important in categorization processes, e.g., in modeling approaches based on evolutionary robotics. Mirolli et al. (2010) present an artificial vision system (composed of fovea and periphery, with simple image processing) that demonstrates the ability to categorise five different kinds of images (letters) of different sizes by exploiting its sensory-motor interactions with its (visual) environment. Similarly, Morlino et al. (2011) demonstrate how a simulated neuro-robot situated in an environment containing parallelepiped objects that (continuously) vary in shape, size, and orientation can develop an ability to associate sensory-motor stimuli with abstract categories and to generalize to new objects. Lanihun et al. (2015) extend the work of Mirolli et al. (2010) by using a more complex image preprocessing technique (HOG) that help to translate to motor responses enhancing the categorization capability for robotic vision control system in the iCub.

Aside from cognitive robotics, ANNs have proven to be useful in various image classification tasks. For instance, Larochelle and Hinton (2010) demonstrated that a Boltzmann machine can be trained to integrate information gathered from several spatially limited glimpses at a static image in order to perform object classification.

Looking at a few other comparable MI models (in terms of using ANNs and their predictive power), some of which are considered to be "representative of the state of the art in the field" (Di Nuovo et al., 2013, p. 217), different approaches can be discerned. These are examined in the discussion.

Our research sets out to accomplish several objectives. Taking the synthetic approach to investigating cognitive phenomena, it is set up as a proof of concept and designed to be exploratory rather than to solve specific problems. Nevertheless, the tasks are set up in a way that conveys the problem-solving capabilities of the model. The main objective of the research is to ground the elusiveness of the phenomenon of MI (see the introductory paragraph) through enactive approaches to vision (as a necessary prerequisite) and enactive approaches to mental imagery. As enactive theories to vision stress the necessity of action for perception, we try to implement this through anticipatory behavior of the model, which needs to make certain movements to get new information and therefore come closer to solving a task. What similar MI models (e.g., Chersi et al., 2013; Seepanomwan et al., 2013; Gaona et al., 2014; Di Nuovo et al., 2011) disregard is the nature of visual construction – new visual information from the environment is gotten not at the same time and in full, which is how computer vision works, but rather sequentially and in limited range, through the use of saccades, while the rest of the experienced rectangular picture is filled-in top-down (see Section 1.1). This is a fundamentally different approach as this is parallel to what happens in MI, with eyes closed.

We try to implement these principles into our model as we see this to be an unused approach and it seems to be fairly more ecological than other similar approaches, making our model more viable and novel. Afterwards, we try to make a bridge from vision to mental imagery, connecting both on the same enactive principles in the same model, making it less phenomenon-specific and more complete in this regard than some similar models (e.g., Mirolli et al., 2010; Morlino et al., 2011, Lanihun et al., 2015). We demonstrate how mental imagery and its use for spatial problem-solving can be grounded in enactive processes (e.g., perceptual actions) as opposed to the grounding of other – arguably competing – theories, especially pictorial and descriptive theories.

As such, our model is a case for enactive approaches to vision and mental imagery, which are still emerging as viable paradigms in empirical, be it analytic or synthetic (Mirolli & Parisi, 2009), research. More generally, our model fits into the paradigms of sensorimotor enactivism and embodied cognition, and therefore lays another piece into the mosaic of the case for their feasibility, especially when the symbolic approaches are still prevalent.

2. The model

2.1. Overview of components

Our proposed model consists of three major modules (components): The first is the forward model (FM) predicting the next state within the configuration of an imagined object, in terms of proprioceptive and categorical information, based on a state and perceptual action input. The second module is the inverse model (IM), which predicts the direction and size of a PA, which can be executed by the robot's visual system. The third module is referred to as memory module (MM) which also has a control function, such that it initiates the FM and IM in one of three tasks (described later) and keeps track of the necessary requirements with memory-like aspects (e.g. the number of executed actions). Furthermore, it serves as a "social" interface between the robot and the task giver.

Figure 4 provides an overview of the proposed system architecture and additionally provides visual information about the most important flow of information within the modules. All three proposed MI tasks can be performed using this configuration, with changes made only in the specification of the MM. Next, we describe the individual modules in more detail.

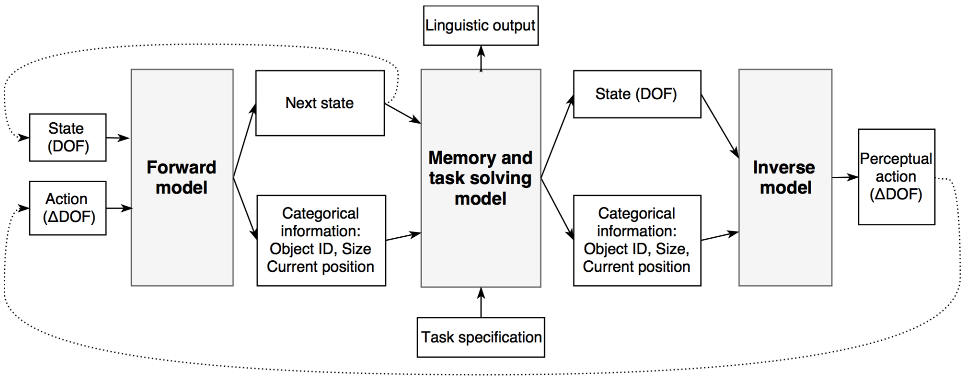


Figure 4: Overview of the proposed system architecture for visuospatial MI. Displayed are the three main components of the system: the forward and inverse models, connected by the controller module. The inputs and outputs for each subsystem are indicated with white boxes. The solid arrows represent the most important information flow necessary for a single PA. The internal update of processed states and performed PAs are indicated with dashed arrows.

2.2. Forward model

The basic idea of a forward model is to predict the next state of the system as a result of an executed action. Our FM, illustrated in Figure 5, has the same function: it takes as input the current state of the positions of the robot's eyes and the action generated by the inverse model and outputs the next state, i.e. the state of the eyes after the executed action. However, our theory grounds concepts in perceptual actions and thus these actions can also answer questions about what the robot is currently looking at. For this reason, our forward model has another use – to recognize the scene. The FM in this work is composed of two neural networks, both fed with the same input (the state and action). One neural network predicts the next state, while the other predicts categorical information about the currently viewed object. This recognition part is not a typical part of the FM, but since PAs are the basis for scene recognition and FM takes actions as inputs, we decided to expand the FM to also act as a scene recognition model.

The categorical information about the scene, provided by the second part of the FM, consists of the current object (in our case a triangle or a square), the object size, direction of the visual trajectory around the object and the current position within the trajectory. Objects and current position have one-hot encoding, direction is binary (0 for counter-clockwise and 1 for clockwise) and size as a continuous[1] value between 0 and 1. There are four possible positions for a square, labeled A–D, and three for a triangle, where D is ignored (see Figure 8).

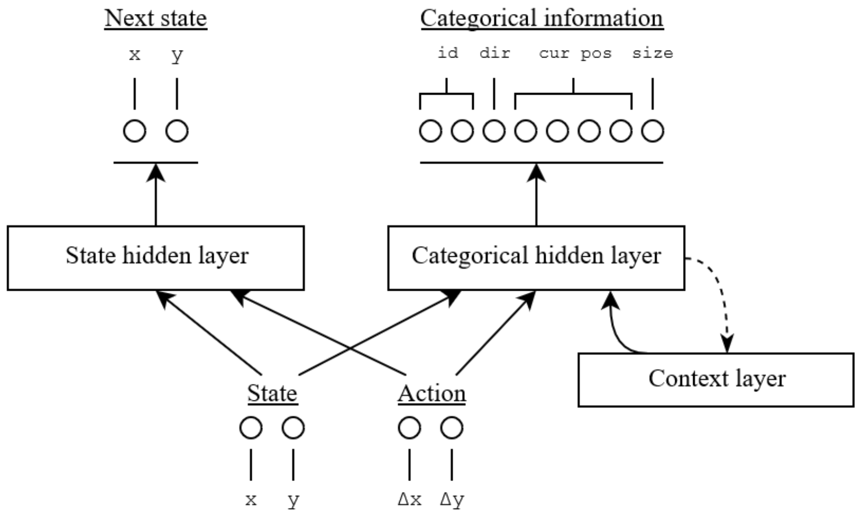


Figure 5: Diagram of the forward model. On its input there are coordinates of the current state (azimuth - x, and elevation - y) and a change of these coordinates as the PA. State hidden layer consists of 17 neurons and the categorical hidden layer contains 45 neurons. The context layer is of the same size as categorical hidden layer. On the output we have 2 coordinates for next state and 8 outputs for categorical information, 2 for one-hot encoding of object ID (10 for triangle, and 01 for square), 1 for a binary direction, 4 for one-hot encoding of the current position and the last one encodes size.

Computing the next state is a trivial operation of adding the action to the state and could be computed directly without the need of a neural network, but because our model is connectionist we decided to use a multilayer perceptron for this computation. Its output is approximate (rather than discrete), making this model closer to biological systems. Because the categorical information can only be extracted from a series of perceptual actions and not from a single one, the categorical part of the FM needs access to previous contexts. This is achieved by using a simple recurrent network (Elman, 1990).

2.3. Inverse model

The overall goal of the IM is to predict the angular values of a single perceptual action. This action is then performed as a saccadic movement by the robot's eye. The eye movement can be expressed in terms of a

[1]More precisely, size is not a categorical information, but for practical reasons we included it here.

vertical and horizontal part (i.e. its elevation and azimuth) so the IM's output consists of two units, each coding for one of them.

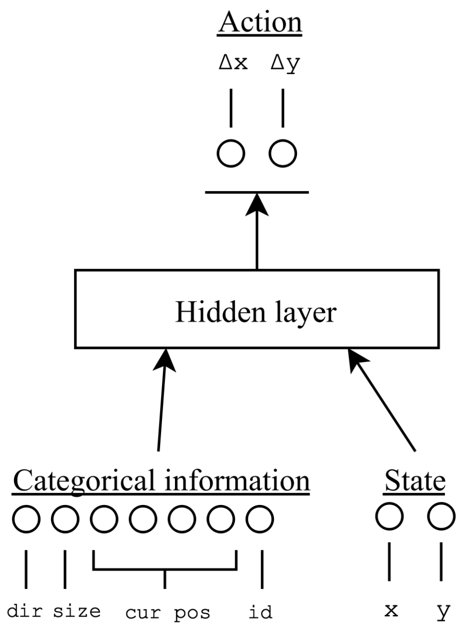


Figure 6: Architecture of the inverse model. Input representation consists of two units encoding the coordinates of the current state (x = azimuth, y = elevation) as well as 7 inputs encoding categorical information. Four units encode the current position in one-hot encoding, one unit represents the size. Two binary units encode the Object ID and processing direction. The output consists of the predicted change in state in azimuth and elevation, i.e. the PAs. The hidden layer has 20 units.

As shown in Figure 6, the IM uses 9 different inputs with an activation range of [0;1]. Two inputs encode the system's current proprioceptive state (azimuth and elevation). Further inputs encode the object ID (one-hot), the parsing direction (0 = counter-clockwise, 1 = clockwise) and the size (continuous). The final four inputs encode the current position within an object (from A to C for triangles and A to D for squares). All values for these input units stem from the predictions of the FM and are therefore based on the overall system's imagined state. No changes were made to outputs of the FM, except for range conversions from [-1;1] to [0;1], if necessary. The two output units, encoding azimuth and elevation of the PA, have an activation range [-1;1]. This range is then transformed into angular values and fed into the robot's gaze controller in order to perform the corresponding PA. The IM has an architecture of a feed-forward network consisting with a single (fully connected) hidden layer of 20 units that connects the described input layer of 9 units with an output layer of two units.

It should be noted that the proposed architecture of the IM differs from a "typical" one presented in other research, as it does not use any target state information as input. Instead, it predicts the PA based only on the current proprioceptive state and the categorical information.

2.4. Memory module

The memory module is implemented as a simple symbol-processing based script, which activates the two remaining networks in order to solve the currently given task. The memory module stores information and provides the required ability to solve the tasks: First, it stores all task-specific variables, such as the type of task currently processed, the object type (triangle or square), as well as required additional information such as size, if needed for a particular task. Second, the MM provides simple verbal feedback in written language to communicate with the human user and indicate the predicted answer to a question. Furthermore, the MM keeps track of the performed PAs and the starting position within the object, and uses this information to decide if a task was solved successfully or not (when the starting position is reached again, the shape trajectory is complete). Its current state is to be seen as a prototype in order to maintain the validity of the model with regard to the underlying theory (more on this topic in the discussion).

The MM calls the two remaining modules repeatedly in order to solve the task at hand. It is responsible for the flow of information from the FM output to the IM input and from the IM output to the FM input. Furthermore, it controls the transformations between azimuth and elevation angle based coordinates that are required as activation values for the robot's visual system and the network's internal activation system.

2.5. Visual processing interface

The described model architecture receives inputs from and outputs commands to an interface of a simulated iCub robot. This interface provides the network with the current proprioceptive state of the robot's eyes. It should be noted that for the described three tasks we employed only one eye of the robot, resulting in mono vision. While this still outputs sufficiently enough visual information about the presented object, it makes any complex stereo-vision computations (such as eye vergence) unnecessary. However, the interface can in theory easily be extended to perform stereo-vision based processing. Any change in the robot's proprioceptive state (and therefore any changes in visual input) are triggered exclusively by PAs commanded by the inverse model described previously.

The actual movement is performed by the iCub's inverse kinematics module (Roncone et al, 2016). It computes a valid path between a given proprioceptive starting and target state. In this implementation, we fixed all available joints of the iCub robot except for two degrees of freedom in eye azimuth and elevation, resulting in non-ambiguous trajectories required in order to reach a particular state. However, the model can be re-used as-is in combination with an inverse kinematics module computing a trajectory for more movable joints (the representation of proprioceptive state would have to be expanded to include all degrees of freedom).

The visual processing interface additionally comprises a simple corner and edge detection test, based on the OpenCV implementation of Harris corner detection. This corner detection routine was further used to compute the start position within an object after "landing" at a random corner of it based on salience. For this, the central point of gravity of the given 2D shape was calculated based on the spatial relations of the corners.

2.6. Unit activation range conversion

All network input and output units require transformations between the activation ranges (ranging between [0;1] and [−1;+1]) for the network units and the actual angle values (azimuth and elevation) that can be reached by the simulated robot's eyes. Based on empirical tests, we implemented several routines to map elevation angles from [−12;+12] range and azimuth angles from [−35;+35] onto [−1;+1] range. Similar routines provide transformation in opposite direction. It should be noted that both the FM and IM use the same transformation scheme. This enables the model to process only [−1;+1] ranged values internally, without remapping back to initial (physical and perceptual) values.

2.7. Implementation

Both the FM and IM were implemented in Theano, with some routines based on the Lasagne package for simplified neural network construction. All neural network scripts were written in Python, while the controller scripts for the iCub simulator consisted of both scripts in Python and C++. All training and testing steps were executed on notebook CPU.

3. Experiments

3.1. Data acquisition

Just as biological agents have to learn to use their bodies to their full capabilities, so did our model need to train on many examples to achieve optimal performance. These examples were gathered from iCub performing PAs in the simulator with the help of iCub's inverse kinematics gaze controller module (Roncone et al, 2016). We created a square and an equilateral triangle, both with side lengths of 25 cm, and changed the floor, background and all surroundings to a white texture, so that only the object could be visible. Because our model deals with PAs in the form of saccades, we were only interested in 2 degrees of freedom, namely eye version (azimuth) and tilt (elevation), and all other iCub's joints except the eyes were turned off. Eye vergence could not be disabled, but since we only worked with the left camera image it did not matter. Object size was manipulated through the distance from the eyes because our main interest was in the saccades performed and not judging the distance (which would be difficult as there were no reference points in the white surroundings and vergence was ignored).

The training procedure for each object was as follows. First an object was presented in the visual field before iCub's immobile head within its visual field, which was limited to $[-35;+35]$ degrees for the azimuth angle and $[-12;+12]$ for the elevation angle. These constraints were set empirically so as to avoid extreme angles where the gaze controller's performance was not guaranteed. Then a Harris corner detection was performed on the seen image to detect salience points and a saccade movement to the nearest corner was performed. The first saccadic movement was not stored as part of the object trajectory because it only represented attention to the object. The next steps had the same structure: first the salience points were detected, a saccade to the nearest point was performed and finally, in order to avoid flipping back and forth between the same corners, evaluation that the new fixation did not match the one before the last PA was done via comparison of eye states. If the PA was valid (i.e. the eye gaze did visit the next corner), it was stored as part of the object. After the whole object had been traversed, its actions were written to a corpus along with categorical information extracted along the way. The training corpus consisted of 2500 instantiations of objects of both shapes and various sizes, all starting positions and directions at various locations within the visual field. An example of the iCub performing the saccades is in Figure 7.

Categorical information about the scene consists of shape information (number of corners), starting position, direction and object size. Shape information was already known at the point of object creation, while the size and the starting position could only be determined after the first action – attention to the object. At this time the whole object was in view and its size could be determined by calculating the portion of the image that it covered and then scaled to $[0;1]$ range, where 0 represents an invisibly small object and 1 represents the size of the largest instantiation of an object that could be seen. The starting position was determined with the help of other salience points (corners), because their average showed their center of mass and thus indicated where the rest of the object lay, relative to the eye focus. Direction of the trajectory was determined after the second action in a sequence when the positions of the first two fixations were known. Current position was then inferred from the starting position, direction and the number of performed saccades.

3.2. Forward model training

The state predicting part of the FM was trained for 30 epochs over all objects in the training set with learning rate 0.01 and the categorical part was trained for 100 epochs with learning rate 0.008. Both parts also used

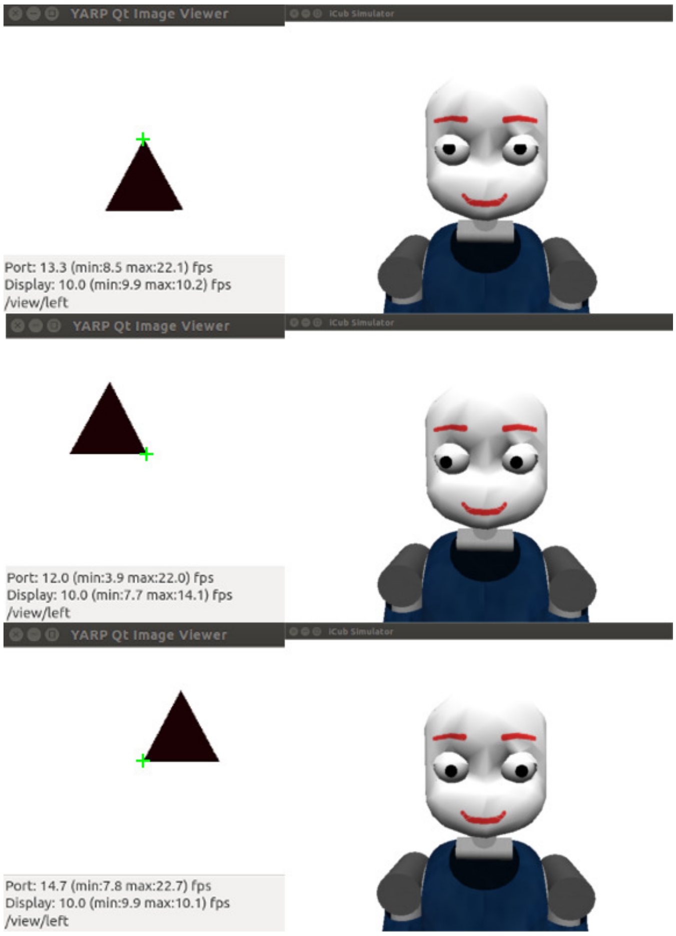


Figure 7: A sequence depicting iCub fixating upon the corners of a triangle. On the left we can see the sequence of images from iCub's eyes, on the right we have the iCub with corresponding eye gaze.

momentum of 0.9 to optimize the learning. Because categorical information differed in how it was encoded – one-hot encoding for object ID and current position, binary for direction and continuous value for size – a bit of tweaking was necessary to optimize the learning of size, because ordinary sigmoid activation resulted in the size neuron always outputting a value very near 0.5. This happened because the (continuous) size information is the noisiest in contrast to other, binary data. For this reason in recognition part we used sigmoid units with a slope $k = 20$ and in the last 25 epochs only size neuron's error was backpropagated constantly while the error from other output units was ignored if it was smaller than 0.1 (in absolute value). In this way the last part of the training was dedicated to fine-tuning the size neuron.

3.3. Inverse model training

The inverse model consists of an input layer spanning 9 units, as explained in Section 2.3. Two input units encode the azimuth and elevation angles of the current state and four input units encode the current position (representing one of the corners for a triangle or a square). Object size, object ID and the processing direction of the object are represented by one input unit each. During processing the values for all input units are generated by the forward model.

The output units have a range of $[-1,1]$ which is transformed directly into the corresponding angular value, as described in Section 2.6. These angular values represent the change in degrees of freedom for azimuth and elevation that can be performed by the iCub robot's gaze module to perform a single PA. Therefore, the angular values for the performed PAs, as retrieved by the iCub's visual interface, can directly be used as targets to train the inverse model.

The output units were equipped with a hyperbolic tangent activation function in order to return values between -1 and 1. The model was trained using stochastic gradient descent with Nesterov momentum by employing the mean squared error between the predicted and target vectors. The training lasted 30 epochs with a learning rate 0.01 and a momentum 0.75.

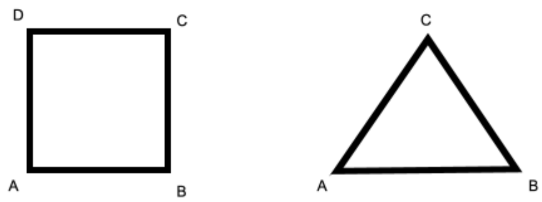


Figure 8: Examples for valid object orientation and corner naming.

Figure 9: Examples of skewed objects located at the edge of vision.

3.4. Simplifications

A variety of simplifications were chosen in order to decrease the task complexity while maintaining its ecological validity: First, the range of imaginable objects is restricted to triangles and squares. Furthermore, triangles are always equilateral and standing up-right. Figure 8 provides an example of two valid objects with corner names. It should be noted that the visual input to the iCub simulator eyes can be skewed significantly, resulting in distorted shapes, i.e. not truly equilateral triangles and curved outlines instead of straight edges; see Figure 9. The presented objects are not rotated, but remain in a fixed orientation, varying only in the location within the robot's visual field and their size. This means that both squares and triangles have a horizontal edge facing downwards (i.e. pointing towards the simulator's ground surface) in the simulator. A further simplification is the aspect of a starting position, for all three tasks the system was trained and tested with the first state within an object.

3.5. Task specifications

Three different tasks have been designed and can be solved by the current implementation of the proposed architecture. Considering the concept of an internal model of the agent (Gigliotta, Pezzulo & Nolfi, 2011), tasks 1 and 2 correspond to an online mode (where the agent receives an input from the environment) and task 3 to an offline mode.

3.5.1. Task 1: Salience-based object recognition

For this task, "What is the input?", the robot's eyes are always open, i.e. visual input is processed for the task continuously. Any performed saccades are salience-driven, leading the robot's eyes around the shape of the presented object. The robot has to predict the identity and the size of the visible object based on 3 (for triangles) or 4 (for squares) saccades. For this task, objects of random size, identity and position were created, constrained to appear within the current field of view. The paths were started at a random corner within the object, and lead in a random direction (either clockwise or counter-clockwise). For more detailed evaluation of the system's performance, predictions were generated after each PA. However, for the final accuracy score, only the final prediction was used, after passing all PAs within the object. The initial saccade towards the object (i.e. the result of the salience of the entire object) was not passed to the system for processing.

3.5.2. Task 2: Imagination-based object recognition

In order to solve this task, "Is this a triangle (a square)?", the robot once again processes visual input with open eyes. However, this time, any performed saccades (i.e. PAs) are purely imagination-driven. Here, the system has to predict size and direction of the PAs required to perform the path of the requested object. The object size is extracted based on salience immediately after "reaching" the object, as described previously. Several simplifications were made for this task. Most importantly, any succesfully reached corner was used to update the system's internal state memory in order to decrease the error generated by multiplying slightly misaligned states, predicted by the forward model. This is in contrast to the imagination task (task 3) and focuses on exactly this aspect of error multiplication within states generated in imagination. For this task, correcting the performed PAs towards any close corner is a valid approach, as saccades in the real world similarly end at points with a certain salience distribution on a local level. Furthermore, this simplification is inspired by microsaccades, as they appear in humans. We suggest that externally correcting the predicted movement resembles micro-saccadic activity to an adequate level. The landing position was checked for each single performed eye motion. If no corner appeared within a fixed range of 30 pixels (i.e. the size of focus or the range of microsaccades), the process was either restarted with the remaining direction or ended if both directions were attempted. As mentioned previously, another simplification was to set the starting state within the object and not allowing for objects within objects. This means that the system always makes a prediction based on a trajectory from the first to the last performed PA. For example, there cannot be four actions of which the last three are a triangle (with valid edges between corners).

This task is more complex problem than the previous one, as now the combined performance of the system is measured. Errors made in the IM can lead to a weaker FM (and thereby combined) performance and vice versa.

3.5.3. Task 3: Object imagination

The third task, "Imagine a triangle (a square)!", requires the robot to output a valid path that corresponds to the given shape identity input. During this process, no visual information from the robot's perception is processed. Therefore, the robot's eyes are closed the entire time. In a purely imaginative process, the system has to predict 3 (triangle) or 4 (square) PAs as well as the corresponding set of 4 or 5 states. Here, the first and the last predicted state should ideally be identical, and the difference between them can be used in order to compute accuracy. The correctness of the path is checked for validity, in terms of a continuous size of PAs and their alignment. The process is instantiated with a randomly generated size value in order to check for prototype effects, i.e. preferred sizes where the combined network operates most efficiently. The generated paths were additionally evaluated by being projected on a flat surface within the field of view and thereby generating a visual trajectory.

Task 3 requires the system to be very accurate in both motor actions as well as in the production of their internal representation. This is mainly due to the fact that, as the task represents a pure imaginative process, the output of the inverse model is not corrected by comparison with an existing visual object. This means that there is significantly more room for error multiplication during object imagination compared to task 2. The output of the categorical part of the FM is used only for validation and is not input into the IM, which receives the task's categorical specifications.

3.6. Memory module in task solving

For task 1, the MM calls iCub's inverse kinematics module and feeds the generated PAs as well as extracted categorical information to the inputs of the FM. This is done for each individual action and followed by a prediction of the system. The prediction is then converted into linguistic labels and printed for accuracy evaluation.

In order to solve task 2, the MM is able to get initialization values from the robot's visual interface and start the task processing by causing the inverse model to generate the first action. This action is then fed into the forward model in order to start the loop that finishes when the last PA is performed. After each performed PA, the MM is used to assess the accuracy of the FM's categorical predictions. In case of mismatch, the loop is discontinued and the task processing is finished if no remaining trajectory directions are left. If the network successfully performs the required amount of PAs, the task is solved. In both cases, the outcome is once again transformed into the corresponding linguistic labels and printed for evaluation.

In task 3, the MM acts as an initializer and the connection between the FM and IM. The initialization occurs as a random choice of starting state, size, direction and starting position. These parameters are input into the IM, which generates the first perceptual action. The FM receives the starting state and first PA to predict the next state. The FM's new state output is then connected to the input of both the IM and FM and the action output of the IM is connected back to the FM's input. Thus a loop is formed which runs until the MM recognizes that the object trajectory is complete. No transformations are needed for the state and action values as this task is processed in unit activation values. The MM additionally checks the FM's categorical output in order to validate that the model is doing correctly. However, only the actual task specifications are input into the IM (i.e. the IM is generating actions for the task at hand.)

4. Results

In this section we describe the performance of the proposed model. The whole corpus obtained with the iCub's inverse kinematics module consisted of 2500 objects which were used both for training and testing of the modules. To test the generalization of the model, we split the corpus into various ratios of train/test data to see whether smaller training set impacts the learning. Ratios tested were 15, 20, 25, 33, 40, 50, 60 and 70% of data used for training, and the final squared test error of the separate and combined models can be seen in Figure 10.

The FM was trained 10 times on each amount of data and the mean error is always around 0.015–0.025, which suggests that the FM generalizes well. We can also observe large deviations at certain points, indicating that the model can get stuck on local minima and better optimization techniques could be used. The combined model was made by taking the best trained model parts and seems to be showing a slow gradual decline, which would indicate that somewhat better learning can be achieved with larger amounts of data. The IM error shows that the model generalizes very well, as there is practically no difference between the errors for the smallest and largest train set; both are around 0.005. In practice, this error translates to around 0.5 degree inaccuracy for both azimuth and elevation angles. Now, we can assess the model performance with respect to three considered tasks.

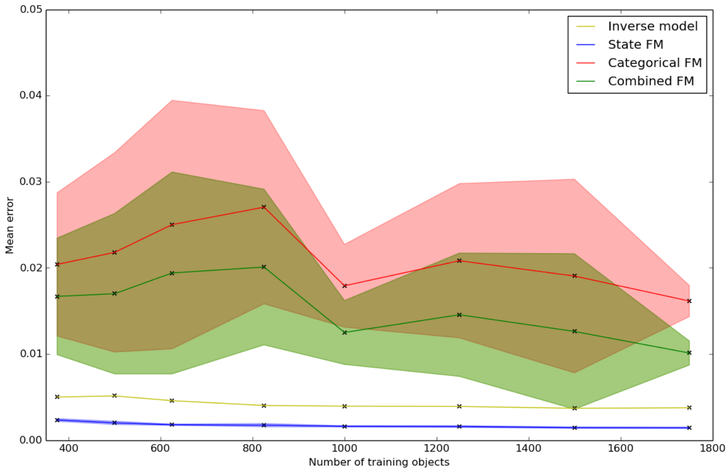


Figure 10: Difference in the final test results of both the FM and IM depending on the amount of training data. The lines denote the mean errors over 10 runs for each training set and the envelopes around the lines represent the standard deviation of the error.

4.1. Task 1: Salience based object recognition

Results for this first task come in the form of accuracy of the prediction of object's shape and size in terms of two linguistic labels for each: triangle/square and small/large. The results are nearly perfect even with the model trained on 625 (representing 25% of the total available training data) objects which proves the generalization ability of the forward model. Both tests were performed for a total amount of 40 objects, split into 20 triangles and 20 squares of various sizes. It should be noted that this task is focused on the forward model accuracy as the action inputs are purely salience-driven and not generated by the system on its own. The errors in the results, which are always regarding the size label, occur entirely at the border region between the two size categories, i.e. where target size is near 0.5 and the model outputs a nearly correct size, but within the wrong category. Figures 11 and 12 show mean size and shape accuracies in the upper two graphs for models trained on 625 and 1750 (representing 25% and 70% of the available data) objects, respectively, and size predictions for both types of objects in the lower two graphs.

The model trained on 625 (25%) objects of the training data reached a mean accuracy of 95% for both triangles and squares for size prediction. The model trained on 1750 (70%) reached 100% for triangles and 95% for squares in size prediction. Both models reached a perfect score of 100% for object identity prediction.

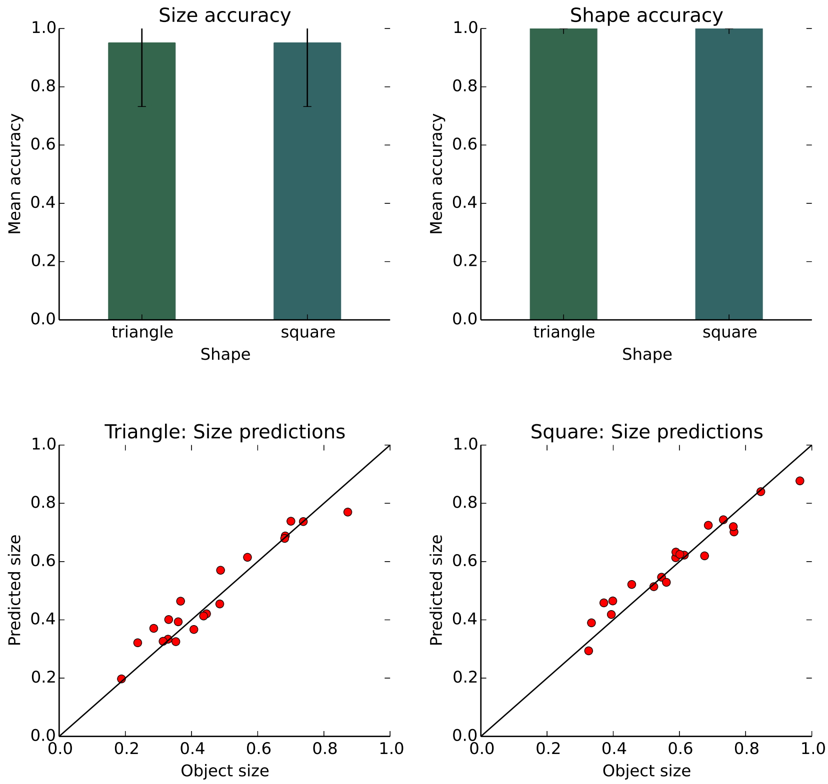


Figure 11: Results of the task 1 of the model trained on 625 objects (25% of the data). Upper graphs depict mean accuracy in size (left) and shape (right) along with standard deviation, lower graphs depict the size predictions (red dots) and targets (black line) for triangle (left) and square (right).

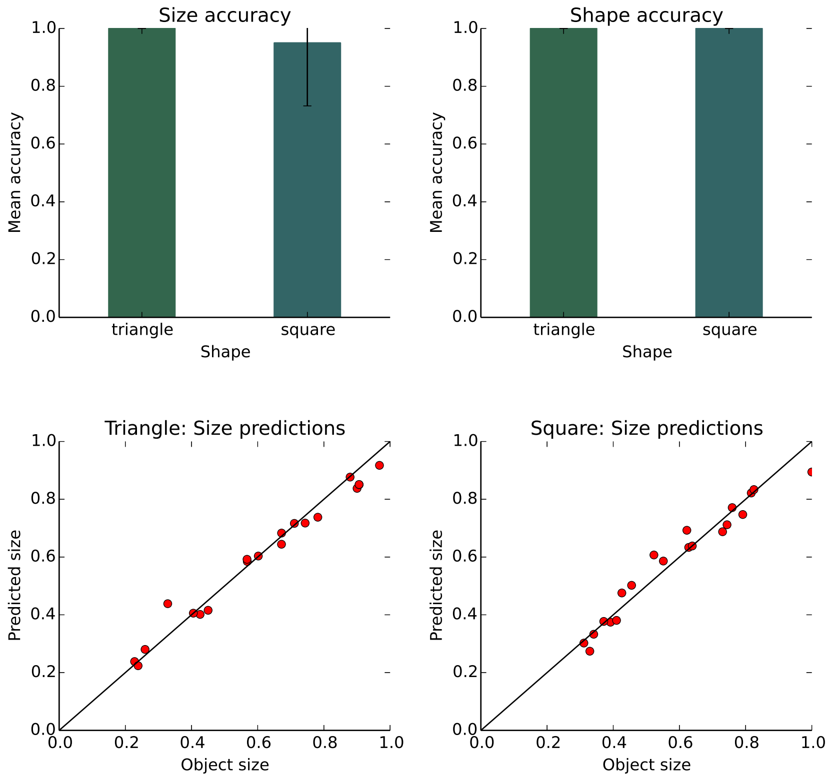


Figure 12: Results of the task 1 of the model trained on 1750 objects (70% of the data). Upper graphs depict mean accuracy in size (left) and shape (right), lower graphs depict the size predictions (red dots) and targets (black line) for triangle (left) and square (right).

4.2. Task 2: Imagination based object recognition

As task 2 was a classification task, we chose to present the results in a confusion matrix, which can be seen in Table 1. The model reached a total score F1 = 0.93 for both object types, for a total number of 80 objects, divided into 40 triangles and 40 squares. The model answered correctly in 94% of the tested examples, with 18 triangles and 17 squares being correctly classified as such and all the incongruent cases (true negatives) recognized. The system failed to recognize 2 triangles and 3 squares and answered that the presented object was not the object in question. The model did not produce any false positives, leading to perfect precision. Additional statistical measures describing the same evaluation, including the reached precision, recall and accuracy are summarized in Table 2. Accuracy accounts to the sum of true positives and true negatives weighted by the total sum of all predicted instances.

Table 1: Confusion matrix of the result for task 2. Cells represent (from top to bottom, left to right): true positives, false negatives, false positives and true negatives.

		answer	
		positive	negative
target	congruent	35	5
	incongruent	0	40

Table 2: Statistical measures obtained from the confusion matrix.

measure	precision	recall	accuracy	F1 score
value	1	0.88	0.94	0.93

4.2.1. Task 3: Object imagination

We present visual trajectories of the imagined objects and measure the accuracies in terms of how close to the starting point the model finished its trajectory of the imagined object. Visual trajectories can be seen in Figures 13 to 16. In each case, the left image displays the trajectory drawn between the imagined states (i.e. where the FM predicted new states based on the IM actions), while the right trajectory shows the perceptual actions performed by the visual system. Each path is a trajectory starting at state 0 and ending at state 3 (for triangles) or 4 (for squares), as a PA connects two neighboring states within a processed object. It should be noted that the model was requested to calculate both the initial state and the final state (i.e. after the last saccade) in order to compute an overall accuracy value for a performed trajectory. Figures 13 and 14 represent two valid instances with good accuracy regarding the initial and end state congruency, while Figures 15 and 16 represent two iterations where the start-end accuracy is worse, resulting in a slightly more deformed shape. Other resulting trajectories are somewhere in between these examples and all of them resemble the ideal shapes quite well. The trajectories based on PAs and the internal states are not exactly the same, due to approximation properties of neural networks.

The results are summarized in Figure 17. The overall mean start-end accuracy for triangles is 96% for azimuth, 88% for elevation and the mean of 92% in both directions. The same variables for squares are 96% for azimuth and 87% for elevation accuracy. The mean accuracy spanning both directions in squares is 91%.

5. Discussion

5.1. Forward model and inverse model performance

The presented preceding tests of standalone forward and inverse model performance can be seen as sanity checks for the combined performance evaluations. Both models reached very good performance in their predefined

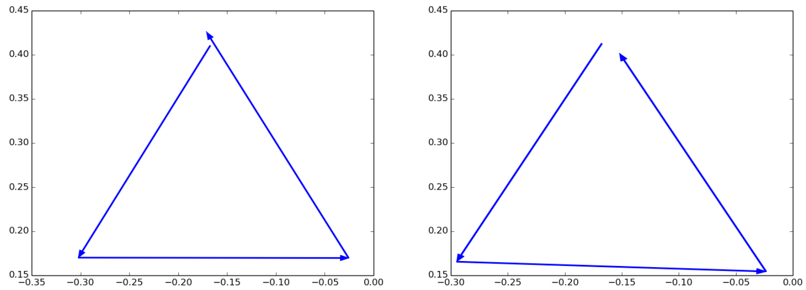


Figure 13: Plots of a nicely performed trajectory for a triangle within the imagination task (task 3). The left image shows the trajectory with states as predicted with the FM, while the right displays the PAs performed by the IM.

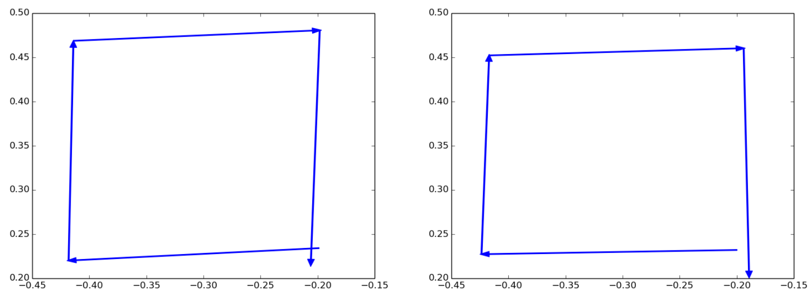


Figure 14: Plots of a nicely performed trajectory for a square within the imagination task (task 3). The left image shows the trajectory with states and the right displays the PAs.

tasks and were evaluated to have the necessary accuracy to be combined into an integrated system. Our main insight during testing the inverse model was that the question how to present the current and previous state is non-trivial and could lead to very different performance. We decided to code the current and previous states in terms of discrete proprioceptive information along with the information about the current position within the object. It should be noted that we tried to keep both models as simple and transparent as possible. This helps with evaluating the combined models' performance and additionally avoids over-fitting the train data set and thereby maintaining the largest possible generalization ability. This is important as the presented objects during test resemble quite strongly those presented in the training data set. As the results of the three tasks show, overfitting the data was not a problem with the presented models.

5.2. Task 1: Salience-based object recognition

Cognitive neuroscience and psychological accounts on salience in humans are neither ubiquitous nor uniformly agreed upon, which means that some

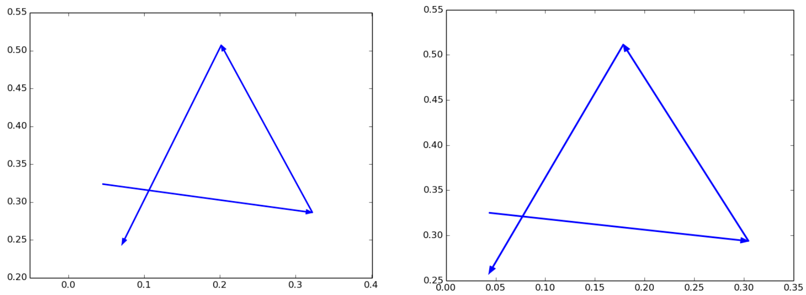


Figure 15: Plots of a performed trajectory for a triangle within the imagination task. The left image shows the trajectory with states and the right displays the PAs. Although the start and end points do not match, the shape is still recognizable.

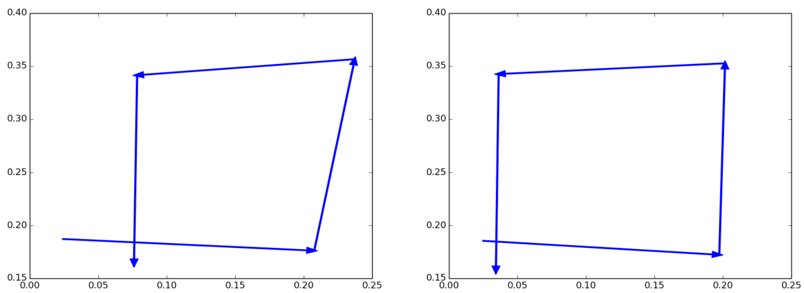


Figure 16: Plots of a performed trajectory for a square within the imagination task. The left image shows the trajectory with states and the right displays the PAs. Although the start and end points do not match, the shape is still recognizable.

parts of our salience-based recognition are not completely ecological. This is especially true when it comes to choosing what is salient for the robot. One of the reasons is that what is salient most probably changes during development, making research on salience very difficult. We settled on looking for corners (in contrast to, for example, colors or edges) not for pragmatic reasons but based on the very simplified "world" that is presented to our robot. Our second choice that was not implemented (partly due to pragmatic reasons) was random eye movement, which might be true in babies. From this, certain patterns and logic may emerge in time, but we decided against such approach. The decision was also due to the fact that our focus did not lie on how salience is learned.

Another phenomenon that we did not tackle is salience in peripheral vision. This is extremely problematic to discuss as peripheral vision itself is such a difficult topic due to how it is (at least partly) constructed in our experience. Salience is therefore even harder to research in peripheral vision. The latter is discussed more in-depth further on.

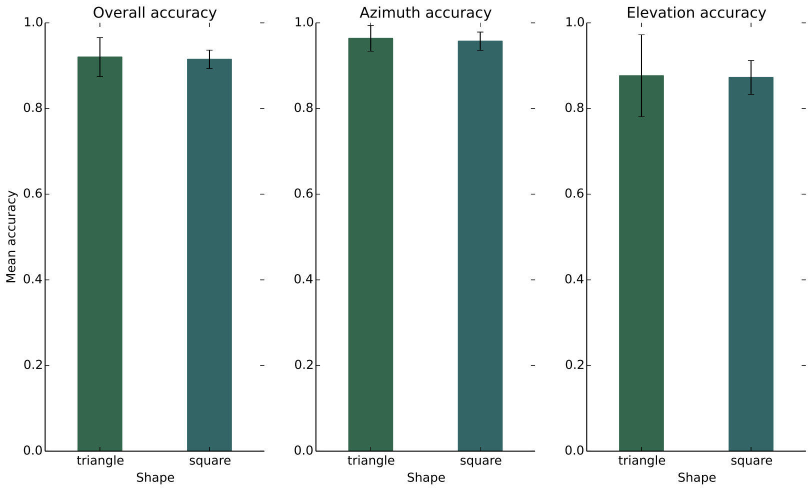


Figure 17: Results of the task 3, depicting the overall accuracy of the final state, and separate dimensions of azimuth and elevation. The error bars depict standard deviation.

When solving this task, the system only processes the PAs performed within the object. This means that the leading and trailing PAs, such as caused by the attention towards the object or shifting away towards the next one, are not handed to the networks for prediction. Solving the task without these artificially introducing breaks between objects can be solved by the model too. In this case, however, further agreements must be found about how to evaluate the predicted identities for trajectories including PAs outside of any objects. One idea is to train the model on an extended dataset which has labels for these actions.

Our model shows very good performance for this task, with mean accuracies ranging from 95% for size recognition to 100% in identity recognition. The difference in overall performance between the data set sizes used for training is minimal and thus the model seems to generalize well even from smaller amounts of training data. The size prediction inaccuracies occurred exclusively when the presented object's size was around 0.5 and consequently it was ambiguous whether this is a small or a large object. However, within an object category, size prediction worked equally well for all presented sizes. This means that the model learned to integrate both the actual size of the object as well as the eventually occurring skewing of saccades due to their nature of being projected on a sphere (in contrast to purely 2D image processing on a flat surface).

5.3. Task 2: Imagination-based object recognition

This task required the model to answer the Yes/No question related to object identity when presented with a single object of either congruent or incongruent identity. Due to the nature of the task and the way the model is trained, no false positive answers were generated. In order to solve this task, the network must (after choosing a processing direction) predict the PAs which would be necessary when looking at the required object (i.e. the object given defined as a linguistic label by the task giver). The task succeeds only if the system is able to accurately locate the corners of the presented object and traverse the path of its edges until all 3 or 4 PAs of the particular object are fulfilled. As we set a fixed focus size of 30 pixels (i.e. the range of simulated microsaccades), this is the accuracy needed to successfully lock to a corner. The decision for a fixed focus area might not be the most accurate and valid decision; a better value might be found in the future work based on research in existing neuroscience literature or with more extensive parameter testing.

The presented model was able to generate 35 true positives and 40 true negative predictions, out of a total of 80 examples. The model parsed objects twice if the ID did not seem to be congruent after the first trial, as there are two possible processing directions for each object's edges. Only five test examples lead to a false negative prediction, when the model was not able to correctly parse and identify the presented object, even though its task description was congruent to the presented visual stimulus.

There is a variety of reasons that can be the cause of errors leading to this misclassification: First, there is still some error for each separate trained network (i.e. the FM and IM), despite using a large dataset of 1750 objects. The recognition of a congruently requested and presented object can fail for two main reasons: Either the IM fails to produce PAs within the focus range or the FM makes an error in classifying the executed actions to be valid. As the results in IM training indicate, there is a remaining error of about half a degree on average in both azimuth and elevation within the IM's predictions. As the learned and performed saccades contain a certain factor of learned skewness, it is non-trivial to differentiate between errors caused by inaccuracy (i.e. miscalculating either the needed size of action or the amount of skewness) or a failed action prediction (e.g. when producing an action fitting a triangle but not a square).

The second possible error source, the FM's performance, can be divided into its required outputs: Either it fails to correctly validate the size or the identity of the currently processed object. It should be noted that the output states are not used within this task and therefore do not influence the system's overall performance. Furthermore, there is the possibility that these inaccuracies could be traced back to the inverse kinematics gaze controller module used to perform the actions in the simulator. The presented implementation has a function to cope with its inaccuracies, but perfect accuracy in motor execution cannot be guaranteed. However, we did not notice this issue as causing the errors within our tests, as the networks' inaccuracies are in general larger.

Yet another possible reason for the described misclassification lies in the empirically defined focus range, i.e. the region that resembles microsaccades in human vision. This region is used to scan for visual corners nearby and correct the performed PA.

5.4. Task 3: Object imagination

The results of task 3 show very good performance for imagined PAs within both triangles and squares. Since this task represents imagination, not perfectly aligned start and end points are to be expected, so long as the trajectory describes a recognizable shape achieved by the model in all iterations. The slight differences in starting and ending location are a result of error multiplication in the feedback loop between the forward and inverse model, since neither is working with mathematically precise values for correct angles and action lengths, but with approximate guesses which are characteristic of natural systems. As the task is specified to be purely imaginative, no external (i.e. visual) correction can be introduced.

One interesting insight that appeared during testing is the fact that the model will generate predictions after each single presented PA. As the presented objects are very simplified, these predictions tended to be correct in all cases before reaching the last saccade. With respect to the underlying theory of PAs, this is a significant aspect: Generating and updating the internal representation of what is processed currently (or rather, what is likely to be processed currently) can be a key decisive factor when choosing the next actions. For the presented tasks within the previous sections, these intermediary predictions are used in a straightforward way, by performing the next PA based on the highest likelihood or by checking if the pattern of highest activation at a given point of time still represents the searched object. With respect to more complex cognitive tasks based on PAs, these intermediary predictions could be exploited in more depth.

In summary, the approach to PAs as the representational medium we chose for presented evaluations is only one possibility. Another approach could be to give the system more freedom for trial-and-error exploration, for example by testing a set of PAs and feeding back positive or negative outcome of a single action, similarly as presented here. However, the system could be re-implemented to perform (multiple) saccades with 'negative' outcome (i.e. not hitting the intended target on the first trial) and performing further saccades from the reached point in space.

5.5. Related neural network models for mental imagery

Here we refer to several connectionist approaches to mental imagery. Chersi et al. (2013) operated with a similar concept to ours when modeling MI. They wanted to exploit its predictive and anticipatory powers, while still relying on comprehensively accurate biological aspects of brain circuits. Their goal was to improve their agent's navigational skills using MI. They designed a computational neural network model of mental simulation where the agent was a virtual rat with several modules: visual areas module, hippocampus module modelled as a self-organizing map, the ventral striatum module working on the method of temporal difference learning, and motor cortex and prefrontal cortex modules which haven't been implemented yet. The MI module was modeled as a multi-layer perceptron. The whole model works as a reinforcement learning model. Mental imagery is used to plot different outcomes for the agent based on its experience. It used the same brain areas used in the actual action performing for imagining, and to a considerable success.

Seepanomwan et al. (2013) used an iCub in their undertaking of an embodied cognitive approach to mental rotation. Their goal was to design successful mental rotating capabilities of their agent. They relied on theories of motor affordance encoding, motor simulation, anticipation of consequences of actions and sensory prediction, which they tried to implement. Their argument is that affordances and embodied processes play an integral role in MI. Their model is composed of four parts: the parietal cortex, receiving proprioceptive and visual input, the premotor cortex, which drives the rotation, the prefrontal cortex, formed by a self-organizing map, which takes outputs of other parts, and the primary motor cortex, which is a self-organizing map as well, encoding current bodily positions and desired or possible bodily positions. The model shows that using the same bodily processes that are used in performed actions can be successfully used in mental rotation.

Gaona et al. (2014) used a ANN model to produce anticipatory behavior using MI. Their goal was to improve their agent's obstacle-avoiding behavior using MI. They used a physical Pioneer 3-DX robot to associate visual and tactile stimuli with prediction, motivated by covert actions. A forward model is used for predictions, as it learns sensorimotor associations from visual, tactile and motor modalities, represented as vectors. Its architecture of a multi-layer perceptron is trained using resilient back propagation to associate environmental stimuli and motor responses. Mental imagery was created by feeding the model its output as input again, building predictive capabilities. The robot was capable of coping with environmental challenges by performing collision-free trajectories. The anticipation of environmental stimuli prepared an appropriate motor response beforehand.

Di Nuovo et al. (2011) used an iCub for modeling spatial MI. Their goal was to build estimative capabilities of the agent through proprioceptive and visual information. Concretely, the model was supposed to imagine scoring a goal, thus improving its performance. The model consists of a fully connected recurrent neural network. Its input is the visual information of the robot's coordinates in respect to the goal and body proprioceptives. The outputs are the desired coordinates and changed body proprioceptives (after performing the action of kicking the ball). The network is trained by Back Propagation Through Time algorithm to predict its own input. The MI serves as a spatial position estimation, based on proprioceptive and visual information. By using MI in such an embodied and predictive way, the model displays successful results.

All these models use the MI for predictive ways, using neural networks to achieve better results. However, our model differs as it uses not only predictive MI, but also predictive visual perception, thus going one step forward from focus on embodiment to focus on sensorimotor enactivism.

5.6. Further work and extensions

5.6.1. Perceptual actions

Since PAs make up a major part of our conceptualization, one of the foremost expansions to be implemented should be to use more of the robot's body. So far, PAs only account for a single eye movement. Going by the enactive theory, the paradigm that different actions extract different information from the environment should be explored further. We limited the amount of possible movements (i.e. degrees of freedom) of the robot to vertical and horizontal eye movements. However, like humans, the iCub platform is able to perform saccades by additionally moving the entire head or can even be supported by moving part of the remaining body, especially the upper torso. Implementing a system that incorporates these additional degrees of freedom would drastically raise its complexity. However, we suggest that our implementation can be seen as a solid foundation for future work on more complex and ecological PAs.

Another missing element in our implementation is the issue of stereo vision. We restricted the perceived visuals to be from a single eye as it still suffices for the range of desired tasks within the scope of this study. However, information received from both eyes is much richer, enabling depth perception (which is exactly the different kind of information you get access to when using different PAs). Implementing a system with stereo vision could significantly help with the problem of skewed saccades, as it is possible to extract more detailed information about the spatial alignment of an edge, or an object as a whole.

Additionally, stereo vision is very likely one of the most important requirements to solve more complex cognitive tasks based on this theory.

5.6.2. Higher level tasks based on perceptual actions

The tasks presented in this study are fairly low-level and are a proof of concept. And, as we explained in the previous sections, already on this level a large number of assumptions and simplifications has to be agreed upon. However, using the system as a foundation for more complex cognitive tasks based on visual PAs is thinkable. These can range from tasks close to what has been described here, such as identifying or imagining objects in the visual field and, for example, detecting the relative positions of multiple objects, their overlap, their relationship in size and so on.

5.6.3. Perceptual actions and supervised learning

Closely connected to the issue of solving higher level tasks using this system is the aspect of how to generate training data, or rather how training should and can be performed within this area of research. More of our assumptions clearly stem from the field of developmental psychology, especially those connected to the problem of salience and peripheral vision. One of the key insights we gained during the study is that there is still no clear definition of what should count as a "correct" PA. Our underlying assumption for training was that PAs should be as precise and efficient as possible, e.g. not checking multiple times for the existence of a corner when solving a task, even though we mostly opted for ecology over optimization. This is clearly reflected in the behavior of the combined model and as well as the evaluation procedure. From a developmental point of view, however, this aspect should be left open to discussion, as PAs, especially during human development, seem to follow not only the rules of accuracy and efficiency, but furthermore support functions such as (random) exploration. One of the most straight-forward tasks that can be implemented is based on the issue of PAs occurring between objects. An extension of the system could be trained to detect them separately to the identity and properties of presented objects.

5.6.4. Memory module

The main reason for adding the MM was to avoid constructing the FM and IM specific to the given tasks. Using a separate MM could enable the combined system to gain the capacity to solve more complex tasks that require even higher-level reasoning. For example, a thinkable task that extends the currently existing framework could be the problem of comparing two visual objects, which are displayed at the same time. Additionally, these objects could exhibit overlapping parts and thus require more complex imaginative reasoning. When relying only on the forward and inverse model to solve this sort of extended task, this would force the task giver to first modify at least one of the models (in this case, most likely the forward model would have to be modified in order to keep track of overlapping objects). Using the memory module, one can distinguish between the "pure" neural organization into a forward and inverse model and a task specific module, which can re-use both networks as they are. From a neuroscience perspective, this resembles the ubiquitous process of re-using and distributing neural activation patterns to form more complex forms of cognitive processing. A very well researched example for this phenomenon would be the scaffolding of the (e.g., human) visual system: Here, "fairly simple" processes such as the recognition of low-level patterns and structures are re-used in large amounts of more complex processes that are additionally supported and guided by top-down processes, for example, when processing a complex object (such as a human) given the task to detect a certain aspect of it (such as a pencil in the human's hand). This is based on the assumption that neural visual processing and imagination modules are not directly affected by the change in the complexity of a visual (or imaginative) task. The memory module in its current state still resembles a simple symbolic controller module. However, in theory it should be possible to design it with a neural network and thereby construct a system based entirely on neural processing. It should be noted that, strictly speaking, the proposed separation into a forward and inverse model is a form of pre-defining the way the tasks are solved on its own.

Within Sima's perceptual instantiation theory, the MM can be seen as an approach to simulate the visuospatial long-term memory (VS-LTM) as well as the short-term memory. Very similarly to the theoretical construct of the VS-LTM, our presented memory module serves as a glue between mental concepts and PAs. It harbors the ability to store and retrieve knowledge for the recognition of a specific entity, such as required for an individual task within our framework. But the memory module also serves within current perception, enabling to identify mental concepts based on PAs and allowing for subsequent interpretation. Within our framework, the interpretation of identified mental concepts is represented implicitly, within the task solving capacity of the module.

The MM probably holds the biggest potential not only for expansion, but also for bringing the model from a one-theory to an extremely versatile theory-testing entity. It was designed so that our model would not be built specifically for certain tasks, therefore implicitly already influencing the research results. However, having the memory module with all the task knowledge (eerily similar to a brain as a central controller) separated from the network (which could be seen analogous to the body) goes thoroughly against the enactive approach, which opposes such dualism or modularism (depending on how the MM is interpreted), making it a double-edged sword. It means that there exists low-level cognition, such as perception, and higher cognition (which would be the memory module). The separation of the two as such therefore goes against the embodied and enactive approaches. We feel it was a necessary compromise in our particular position, but it will be further conceptualized and looked into for options that would be compatible with our paradigm. One of the most straight-forward approaches to this problem could be to train and test the model as a whole, with all parts being a type of artificial neural network. As described before, the memory module used for our tests is only a prototype that can be implemented as a feedforward (or recurrent, for more complex tasks) network. In theory, by connecting the memory module to all input and output units of both forward and inverse model, the system could be trained in a single step.

This is in contrast to our highly modular approach to show the very basic functionality of visuospatial MI in robotic vision. One of the most interesting possibilities of such an integrated approach is the ability of on-line learning, where one part of the network uses the predictions of another one and provides correctional feedback and vice versa.

5.6.5. Inherent properties

Some properties inherent to the module should be discussed as well, especially from the ecological view. Most of the inherent properties not being learned or being presupposed was not only due to the speculative nature of the phenomenon but also because of our focus on the parts necessary for our research and specific tasks. One of such inherent properties is the identity of the object that the model automatically possesses. This would definitely have to be reassessed in the potential expansions, especially when more objects are to be presented in space and time, meaning that the model would have to learn how to differentiate as well as to know which object was already presented to it. Another thing are the constraints in the iCub's visual field. These were empirically and pragmatically set so as to avoid extreme angles, where its successful performance was not guaranteed. Another, probably unavoidable property at this stage, is the categorical information about the scene. Getting the shape information at the time of object creation might be similar to some sort of external linguistic signal, but this still seems oversimplified from the developmental point of view.

5.6.6. Environment complexity

Tackling the task of object recognition with PAs in the real world (as opposed to the presented tasks within the simplified virtual environment) will bring significant additional complexity to the suggested theory and implementation. Most importantly, extracting salient information (or structured information at all) becomes a highly ambiguous task. It is an open question, how and what exactly we extract from our environment, and how we orient in the enormous amounts of salient visual inputs. Combining the proposed implementation of attention-based and narrow visual focus with peripheral vision or even approaches from computer vision could help to tackle this problem.

5.6.7. Peripheral vision

As can be discerned from the constructive powers of visual perception, peripheral vision is a very difficult topic to tackle as it is hard to say how much of it is constructed and how much of this construction is bottom-up as opposed to top-down. It seems that some information must come through as the salience-based vision reacts to stimuli in the peripheral vision. The role of saccades in this is uncertain as well, but it seems that it must be connected to it. One of the possible mechanisms might be occasional saccades to periphery to keep track of significant changes and to pick up on salience. These saccades, however, would not be salience-based. In any case, different approaches to research peripheral vision seem to be perfect for our expanded model which could work as a theory-tester. It is definitely a fascinating research path to be undertaken, and one that might be a potential future path for our model.

6. Conclusion

We presented a simplified simulation of visuospatial mental imagery based on the Perceptual Instantiation Theory (PIT), as presented by Sima, through the larger context of the enactive approach to visual perception, and with added constructing saccadic visual phenomena as a novel inclusion, especially in relation to robotic vision. The theory is built around the core assumption that in humans, PAs are used within perception of the environment in order to extract information and can be "re-used" in MI later on. Our model proves that it is possible to ground simple mental concepts, namely triangles and squares, only on PAs expressed by eye movements.

For this, we propose a system consisting of two artificial neural networks, containing an inverse model, which predicts the coordinates change for a single PA based on categorical information about the imagined object as well as information about the current proprioceptive state, and a forward model, which cares about the internal representation of the current state and furthermore enables object recognition by predicting categorical information based on previously performed PAs. The presented inverse and forward models are connected by a memory module. This memory module was implemented as a simple symbolics controller, which mediates task-specific information to the two neural networks and initiates object recognition and imagination.

The presented artificial neural system works in real-time within a simulated iCub cognitive humanoid robotic platform, using only its eye movements as possible degrees of freedom for PAs. This setup enables training and evaluation with PAs, extracted directly from the eye movements performed by the robot after being presented with objects of variable shape, size and location. In its current state, the system is able to recognize presented visual objects of two shapes (triangle and square) for continuous sizes and locations within the field of view.

The combined model proved to be efficient in evaluating the congruency of presented objects and given object identity labels. Furthermore, the system was successful in imagining valid trajectories of the discussed object types. Overall, only minor inaccuracies appeared within object imagination, which can be traced back on the high variability within possible object configurations within training and testing and the inherent continuous approximation properties of neural networks. Furthermore, the presented system showed the ability to generalize upon the skewness of objects when they were presented at the border areas of the robot's field of view.

The system should be seen as a proof-of concept implementation of the PIT, complemented with the larger context of enactive approaches, research-backed task picks and a novel inclusion of the saccadic phenomena in relation to visual construction. It could serve as a platform to test more extensive simulations based on these. One possible starting point for this could be the presented symbolic memory module, which could be replaced by an artificial neural network, making the entire system connectionist. We suggest that such an extended version of the presented system would provide the possibility to significantly scale up the complexity of solvable tasks and testable research hypotheses.

References

Arbib, M. A. (1981). Perceptual structures and distributed motor control. In Brooks V.B. (Ed.), *Handbook of Physiology: The Nervous System II. Motor Control* (pp. 1449–1480). Bethesda, MD: American Physiological Society.

Barsalou, L. W. (1999). Perceptual symbol systems. *Behavioral and Brain Sciences, 22*(4), 577–660.

Bartolomeo, P., & Chokron, S. (2002). Orienting of attention in left unilateral neglect. *Neuroscience and Biobehavioral Reviews, 26*(2), 217–234.

Bays, P. M., & Husain, M. (2007). Spatial remapping of the visual world across saccades. *Neuroreport, 18*(12), 1207–1213.

Blackmore, S. J., Brelstaff, G., Nelson, K., & Troscianko, T. (1995). Is the richness of our visual world an illusion? Trans-saccadic memory for complex scenes. *Perception, 24*(9), 1075–1081.

Block, N. (1981). *Imagery*. Cambridge, MA: MIT Press.

Bourlon, C., Oliviero, B., Wattiez, N., Pouget, P., & Bartolomeo, P. (2011). Visual mental imagery: What the head's eye tells the mind's eye. *Brain Research, 1367*, 287–297.

Chambers, D., & Reisberg, D. (1985). Can mental images be ambiguous? *Journal of Experimental Psychology: Human Perception and Performance, 11*(3), 317–328.

Chersi, F., Donnarumma, F., & Pezzulo, G. (2013). Mental imagery in the navigation domain: a computational model of sensory-motor simulation mechanisms. *Adaptive Behavior, 21*(4), 251–262.

Di Nuovo, A., De La Cruz, V. M., & Marocco, D. (2013). Special issue on artificial mental imagery in cognitive systems and robotics. *Adaptive Behavior, 21*(4), 217–221.

Di Nuovo, A. G., Marocco, D., Di Nuovo, S., & Cangelosi, A. (2011). A neural network model for spatial mental imagery investigation: A study with the humanoid robot platform iCub. In *Proceedings of 2011 International Joint Conference on Neural Networks (IJCNN)* (pp. 2199–2204), San Jose, CA, USA.

Findlay, J. M. (1982). Global visual processing for saccadic eye movements. *Vision Research, 22*(8), 1033–1045.

Eysenck, M. W. (2012). *Fundamentals of cognition.* New York, NY: Psychology Press.

Fodor, J. A. (1975). *The Language of Thought.* New York, NY: Thomas Crowell.

Gaona, W., Escobar, E., Hermosillo, J., & Lara, B. (2014). Anticipation by multi-modal association through an artificial mental imagery process. *Connection Science, 27*(1), 1–21.

Gibson, J. J. (1986). *The Ecological Approach to Visual Perception.* Psychology Press.

Gigliotta, O., Pezzulo, G., & Nolfi, S. (2011). Evolution of a predictive internal model in an embodied and situated agent. *Theory in Biosciences, 130*(4), 259–276.

Hancock, S., Gareze, L., Findlay, J. M., & Andrews, T. J. (2012). Temporal patterns of saccadic eye movements predict individual variation in alternation rate during binocular rivalry. *i-Perception, 3*(1), 88–96.

Hooge, I. T. C., & Frens, M. A. (2000). Inhibition of saccade return (ISR): Spatio-temporal properties of saccade programming. *Vision Research, 40*, 3415–3426.

Humphrey, G. K., Symons, L. A., Herbert, A. M., & Goodale, M. A. (1996). A neurological dissociation between shape from shading and shape from edges. *Behavioural Brain Research, 76*(1–2), 117–25.

Kosslyn, S. M. (1980). *Image and Mind.* Cambridge, MA: Harvard University Press.

Kosslyn, S. M. (1994). *Image and Brain: The Resolution of the Imagery Debate.* Cambridge, MA: MIT Press.

Laeng, B., & Teodorescu, D-S. (2002). Eye scanpaths during visual imagery reenact those of perception of the same visual scene. *Cognitive Science, 26*, 207–231.

Lanihun, O., Tiddeman, B., Tuci, E., & Shaw, P. (2015). Improving active vision system categorization capability through histogram of oriented gradients. In C. Dixon & K. Tuyls (Eds.), *Towards Autonomous Robotic Systems* (pp. 143–148). Cham: Springer International Publishing.

Larochelle, H., & Hinton, G.E. (2010). Learning to combine foveal glimpses with a third-order Boltzmann machine. *Advances in Neural Information Processing Systems, 23*, 1243–1251.

Lathrop, S. D. & Laird, J.E. (2009). Extending cognitive architectures with mental imagery. In *Proceedings of the 2nd conference on Artificial General Intelligence* (pp. 97–102). Amsterdam: Atlantis Press.

McMahon, C. E. (1973). Images as Motives and Motivators: A Historical Perspective. *American Journal of Psychology, 86*(3), 465–90.

Mirolli, M., & Parisi, D. (2009). Towards a Vygotskyan cognitive robotics: The role of language as a cognitive tool. *New Ideas in Psychology, 29*(3).

Neisser, U. (1976). *Cognition and Reality.* San Francisco: W. H. Freeman.

O'Regan, J. K., & Noë, A. (2001). A sensorimotor account of vision and visual consciousness. *Behavioral and Brain Sciences, 24*(5), 939–1031.

O'Reilly, R. C., & Munakata, Y. (2000). *Computational Explorations in Cognitive Neuroscience: Understanding the Mind by Simulating the Brain.* Cambridge, MA: MIT Press.

Paivio, A. (1986). *Mental Representations: A Dual Coding Approach.* New York, NY: Oxford University Press.

Palmiero, M., Piccardi, L., Nori, R., Palermo, L., Salvi, C., & Guariglia, C. (2016). Editorial: Creativity and Mental Imagery. *Frontiers in Psychology, 7*, 1280.

Pylyshyn, Z. W. (1973). What the mind's eye tells the mind's brain: A critique of mental imagery. *Psychological Bulletin, 80*(1), 1–25.

Pylyshyn, Z. W. (2002). Mental Imagery: In search of a theory. *Behavioral and Brain Sciences, 25*(2), 157–182.

Roncone, A., Pattacini, U., Metta, G., & Natale, L. (2016). A Cartesian 6-DoF gaze controller for humanoid robots. In *Proceedings of Robotics: Science and Systems.* AnnArbor, Michigan.

Seepanomwan, K., Caligiore, D., Baldassarre, G., & Cangelosi, A. (2013). Modelling mental rotation in cognitive robots. *Adaptive Behavior, 21*(4), 299–312.

Sima, J. F. (2014). *A Computational Theory of visuospatial Mental Imagery* (Unpublished dissertation). The University of Bremen, Germany. Retrieved September 25, 2016, from http://cosy.informatik.uni-bremen.de/sites/cosy/files/sima/thesis_imagery.pdf

Szeliski, R. (2011). *Computer Vision: Algorithms and Applications.* London: Springer-Verlag London.

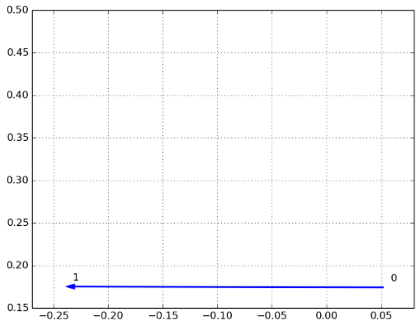
Thomas, N. J. T. (1999). Are theories of imagery theories of imagination? An active perception approach to conscious mental content. *Cognitive Science, 23*(2), 207–45.

Thompson, E. (2007). Look again: Phenomenology and mental imagery. *Phenomenology and the Cognitive Sciences, 6*(1–2), 137–170.

Appendix

In Figures 18 to 21 we present a step-by-step processing of the system in the task 3 with the model imagining a large square (object ID = [0,1], size = 0.967), starting in B position ([0,1,0,0]) and going in clockwise direction

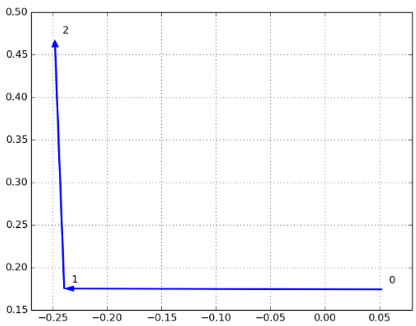
(direction = 1). The system was initiated in a random start state [0.133, −0.744]. Note that the values of the system are in unit activations [−1, 1], while values on the graphs are in iCub's world coordinates.




```
----->Step 1
Prdicting first action with inverse model.
    IM input:   [0.133, -0.744]             # STATE 0
                1.0                         # OBJECT ID
                1.0                         # DIRECTION
                0.967                       # SIZE
                [0.0, 1.0, 0.0, 0.0]        # POSITION
    ================================================================
    IM output:  [-0.805 -0.011]             # ACTION 1

Predicting next state with forward model.
    FM inputs:  [0.133, -0.744]             # STATE 0
                [-0.805 -0.011]             # ACTION 1
    ================================================================
    FM output:  [-0.343, -0.733]            # STATE 1
                [0.594, 0.391]              # OBJECT ID
                0.98                        # DIRECTION
                [0.998, 0.0, 0.001, 0.001]  # POSITION
                0.639                       # SIZE
```

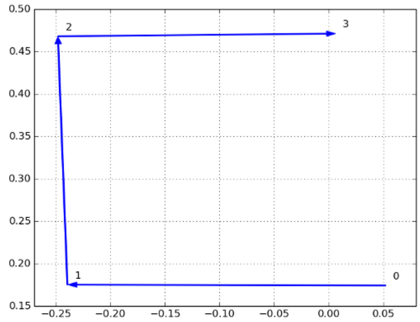
Figure 18: *Left:* The start state and the state after the first step, connected with the starting perceptual action. *Right:* Output of the system in step 1.




```
----->Step 2
Predicting the action corresponding to FM output with inverse model.
    IM input:   [-0.343, -0.733]            # STATE 1
                1.0                         # OBJECT ID
                1.0                         # DIRECTION
                0.967                       # SIZE
                [1.0, 0.0, 0.0, 0.0]        # POSITION
    ================================================================
    IM output:  [0.047, 0.832]              # ACTION 2

Predicting next state with forward model.
    FM inputs:  [-0.343, -0.733]            # STATE 1
                [0.047, 0.832]              # ACTION 2
    ================================================================
    FM output:  [-0.353, 0.567]             # STATE 2
                [0.055, 0.968]              # OBJECT ID
                1.0                         # DIRECTION
                [0.0, 0.0, 0.042, 0.965]    # POSITION
                0.814                       # SIZE
```

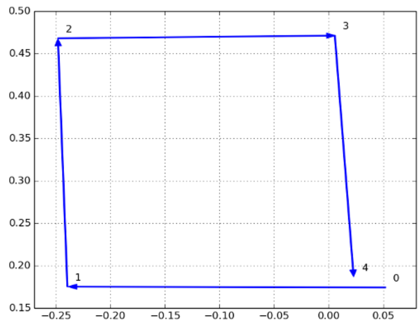
Figure 19: *Left:* The states up until and including the second step as well as the perceptual actions connecting them. *Right:* Output of the system in step 2.




```
----->Step 3
Predicting the action corresponding to FM output with inverse model.
    IM input:   [-0.353, 0.567]              # STATE 2
                1.0                          # OBJECT ID
                1.0                          # DIRECTION
                0.967                        # SIZE
                [0.0, 0.0, 0.0, 1.0]         # POSITION
    ================================================================
    IM output:  [0.779, -0.023]              # ACTION 3

Predicting next state with forward model.
    FM inputs:  [-0.353, 0.567]              # STATE 2
                [0.779, -0.023]              # ACTION 3
    ================================================================
    FM output:  [0.059, 0.575]               # STATE 3
                [0.016, 0.98]                # OBJECT ID
                1.0                          # DIRECTION
                [0.0, 0.074, 1.0, 0.0]       # POSITION
                0.858                        # SIZE
```

Figure 20: *Left:* The states up until and including the third step as well as the perceptual actions connecting them. *Right:* Output of the system in step 3.




```
----->Step 4
Predicting the action corresponding to FM output with inverse model.
    IM input:   [0.059, 0.575]               # STATE 3
                1.0                          # OBJECT ID
                1.0                          # DIRECTION
                0.967                        # SIZE
                [0.0, 0.0, 1.0, 0.0]         # POSITION
    ================================================================
    IM output:  [0.012, -0.830]              # ACTION 4

Predicting next state with forward model.
    FM inputs:  [0.059, 0.575]               # STATE 3
                [0.012, -0.830]              # ACTION 4
    ================================================================
    FM output:  [0.085, -0.692]              # STATE 4
                [0.0, 1.0]                   # OBJECT ID
                0.999                        # DIRECTION
                [0.017, 1.0, 0.0, 0.0]       # POSITION
                0.849                        # SIZE
```

Figure 21: *Left:* The states up until and including the last step as well as the perceptual actions connecting them. *Right:* Output of the system in step 4. As the shape is now complete (position is the same as the start position), the task is now finished.