Robot Multi-Modal Object Perception and Recognition: Synthetic Maturation of Sensorimotor Learning in Embodied Systems

Raphaël Braud, Alexandros Giagkos*, Patricia Shaw, Mark Lee and Qiang Shen

***Abstract*—It is known that during early infancy, humans experience many physical and cognitive changes that shape their learning and refine their understanding of objects in the world. With the extended arm being one of the very first objects they familiarise, infants undergo a series of developmental stages that progressively facilitate physical interactions, enrich sensory information and develop the skills to learn and recognise. Drawing inspiration from infancy, this study deals with the modelling of an open-ended learning mechanism for embodied agents that considers the cumulative and increasing complexity of physical interactions with the world. The proposed system achieves object perception, and recognition as the agent (i.e., a humanoid robot) matures, experiences changes to its visual capabilities, develops sensorimotor control, and interacts with objects within its reach. The reported findings demonstrate the critical role of developing vision on the effectiveness of object learning and recognition and the importance of reaching and grasping in solving visually elicited ambiguities. Impediments caused by the interdependency of parallel components responsible for the agent's physical and cognitive functionalities are exposed, demonstrating an interesting phase transition in utilising object perceptions for recognition.**

***Index Terms*—Multi-modal object learning, vision, reaching, developmental learning, longitudinal study, iCub robot**

I. Introduction

Humans are capable of recognising, classifying and manipulating objects even without prior knowledge or memory about them. This allows successful interactions and the ability to utilise any particular object instance quickly. Acquiring the necessary skills to achieve such a high-level set of cognitive functions starts in early infancy. Developing babies progressively build an understanding of their embodied selves in relation to the world and the objects in it. This understanding emerges from the ongoing exercise of motor activities and the integration of any associated sensory information. As a result, multi-modal perceptions of physical objects are conceived and recalled when necessary.

One of the first objects an infant experiences is its own extended arm, resulting from the asymmetrical tonic neck reflex [1]. Soon the infant begins to discover not only how to perceive the new object as it moves visually, but also how to gradually take control of its movement and interact with other objects. This process depends on the parallel and interdependent development of physical and cognitive capabilities as the infant undergoes several maturational changes. Inspiration drawn from the developing infant is one of the driving forces for designing robotic systems. Robots that can autonomously elaborate sensorimotor dependencies, and fruitfully exploit multi-modal sensory information while acting, are anticipated to address everyday challenges in dynamic environments.

This work describes a robotic system architecture that demonstrates the longitudinal development of embodied agents with respect to open-ended learning of object perception. A humanoid robot is presented with a number of small objects, including its hand, and builds multi-modal perceptions while undergoing developmentally plausible changes.

The longitudinal development followed in this work builds on previous work found in [2]. At the end of every stage, the system utilises what is learned in a multi-modal fashion, to recognise the objects. The results demonstrate the efficacy of autonomous learning and its impact on reducing uncertainty and visual ambiguities via active interactions as opposed to passive observations [3].

The rest of this paper is structured as follows. Section II introduces the architecture of the system and discusses the participating modules in connection to their role in development. In section III, the processes of multi-modal learning and recognition are presented. The experimental methodology is given in section IV. The findings of the experiments are documented and discussed in section V. Finally, conclusions are available in section VI.

II. System Architecture

The proposed system consists of several interconnected and interdependent modules, as shown in Figure 1. Each module plays a distinct and important role. Note that since an iCub robot [4] is used throughout this study, the body maps discussed here reflect that particular humanoid's joint structures. The following modules provide the main algorithmic infrastructure. [1]

A. Longitudinal monitoring and control

The system employs the Lift-Constraint, Act and Saturate (LCAS), a stage-based developmental algorithm for building robot controllers inspired by developmental psychology. LCAS facilitates the progressive learning of sensorimotor and cognitive capabilities, based on the idea that constraints that represent anatomical and maturational impedances during

Department of Computer Science, Aberystwyth University, Wales, UK
*Corresponding Author: alexandros@giagkos.com

[1] A list of videos that demonstrate the system's capabilities is found at https://www.youtube.com/playlist?list=PLhcQ58f13VtM42rIIVYeUIROam6zIxXew

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TCDS.2020.2965985, IEEE Transactions on Cognitive and Developmental Systems

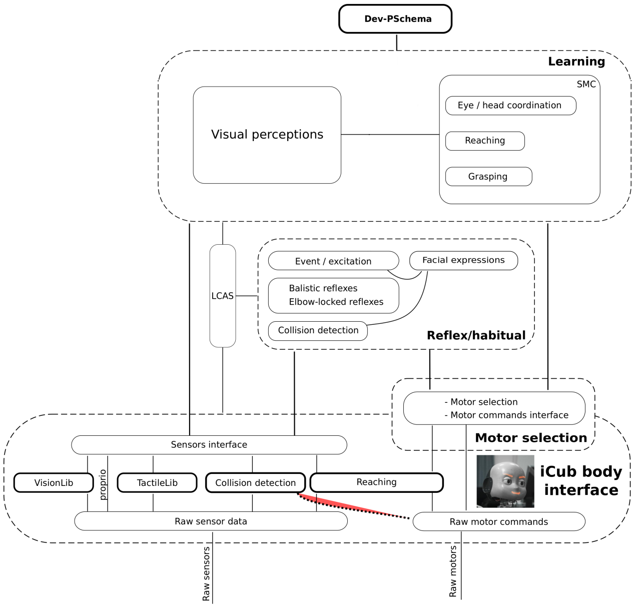


Fig. 1: The system architecture. The modules are discussed in section II.

infancy are gradually lifted. To lift a constraint, and thus mark the next stage in the development, learning has to have reached saturation, i.e., no further progress is possible. Detailed descriptions and applications of LCAS in robotic development are found in [5], [6]. As seen in Figure 1, LCAS is connected to all other components to monitor the saturation levels and to coordinate the progression of the development.

B. Maps and receptive fields

Drawing inspiration from the way most areas of the cortex are organised, i.e., in two-dimensional topographical layers [7], the main data structures are multi-dimensional arrays, hereafter referred to as maps, and are used to represent both sensory and motor spaces. Maps are not only used to store values; they also constitute a major mechanism by which learning is achieved. For instance, during sensorimotor activity, the sensory and motor values are used to activate small regions of the corresponding maps and to learn explicit links between them. These small regions are named fields by analogy with the receptive fields in the brain, and consist of a centre point and an activation radius. Consequently, repeated associations drawn between activated fields accomplish learning in a Hebbian fashion. More information about maps and their efficacy in learning can be found in [8]. In this work, the use of maps and fields to accommodate data related to visual perception and model recognitions is demonstrated.

C. Vision and feature detection

The vision module is located at the lower part of the architecture, denoted as *"VisionLib"*, and is designed to provide two main functionalities. Firstly, it alters the field of view (FOV) and acuity of the two camera images to emulate infant vision according to the developmental time-line derived from the psychological literature [9]–[14]. The effect of applying those vision alterations is depicted in Figure 2. Secondly, the module applies four low-level feature extraction techniques that identify particular features in the images, namely colour, brightness, edges and motion.

In brief, feature detection is employed to identify the retina coordinates of any stimulating coloured target, whose region's average colour is within a pre-defined Hue, Saturation and Value (HSV) range of values. HSV is preferred over other colour models such as RGB due to its robustness towards external lighting changes, with Hue varying relatively less in real-world environments. Similarly, salient targets owing to their level of brightness as measured by Value are extracted, with their coordinates, as well as sizes, being also returned. Note that the size of these regions (in pixels) is measured to make colour and brightness distinguishable targets, identified at different locations in the retina.

Furthermore, the Canny edge detection algorithm [15] is used to identify targets that are defined by the same convex shape. Once identified, their perimeters and the Euclidean distances between their two extreme horizontal and vertical points are measured. These three numbers are then used to both characterise and distinguish stimulating targets based on their shape information. Finally, motion detection compares two consecutive images, identifies the coordinates and measures the sizes of moving targets in the retina. Note that size thresholds are used to minimise the noise in the extraction of all features. A more detailed description of the vision module is found in [16], [17].

D. Learning saccades

The system learns to control its gaze between identified targets because of the efforts of the module depicted as *"Eye/head coordination"* in Figure 1. Learning algorithms that utilise the extracted features in the retina, with respect to the available vision capabilities at each stage of development, populate maps related to the sensorimotor control of the eyes and neck. A detailed description of these algorithms is given in [18]. In brief, by performing eye and neck motor babbling, the system aims at fixating an interesting stimulus. When a target is centred, the relative motor values responsible for the successful saccade are used to activate the corresponding receptive fields for the eyes and the neck. In retina map is defined by the height and the width in the input image, whereas the motor maps are defined by the pan and tilt joints of the eyes and neck, respectively. Activated fields are in turn linked to allow future saccades between learned areas in the retina.

Given sufficient time, the agent develops the ability to bring any desired area to the retina's centre, and thus to perform precise saccades. Note that due to the longitudinal approach and in particular the progressive improvements to vision, the learned body maps grow in population, making the system's ability to saccade more efficient at later stages.

E. Learning reaching by hand regard

Humans undergo a long period of self-exploratory engagement that plays a key role in discovering their embodied self

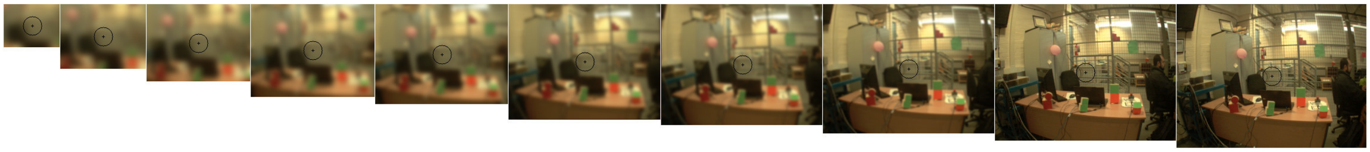


Fig. 2: The result of altering the FOV and acuity to images in order to reflect the vision capabilities of an infant per developmental stage (1–10).

and in understanding their own body as a unique entity in the environment [19]–[21]. Starting from the second month, a hand regard behaviour is observed as the hand attracts much attention, while it manoeuvres within the infant's field of view. This behaviour becomes less common after the fourth month when the control of the hand has progressed. From this month onwards, when the visual stimulus of an object attracts its attention, the infant can occasionally bring the hand slowly to the object position, while saccading between the hand and the object [22].

TABLE I: List of symbols.

Symbol	definition
$v \in V$	visual field v in the gaze space V
$m \in M$	motor field m in a motor map, e.g., left arm's
$k, t \in K$	position fields in iCub's ego-centric 3D space
K_{pop}, K_{stop}	size and threshold for map population
vel	next set of velocities to be sent to modality
T	vector of directional information to target in K
E^A, E^M, E^S, E^D	heuristic events for PO creation
S^i, S^p	strengths of confidence for instances and pairs
N^c, N^o, N^m	number of common, observed and memory features
M^o, M^m, M	scores for similarity: observed, memory and overall
PO, G^{PO}	proto-object and proto-object graph
p_x, cPO_x	ambitious PO candidate x and associated confidence

Similarly to saccading, moving the hands within the reachable space is a result of a staged learning process during which the associations between sensory and motor values are discovered. Drawing inspiration from the hand regard behaviour described before, the module denoted as *"Reaching"* in Figure 1 is responsible for building and linking together gaze-related and motor-related maps associated with reaching control.

Our ability to perceive the three-dimensional world starts to develop in early infancy, as we make use of a number of kinetic, monocular and binocular cues. Depth information is a gradually refined process. At month one, infants only focus on objects 20 to 25 centimetres from their faces and are mainly attracted by objects in motion whose spatial information can be easily detected [23]. Although infants display sensitivity to depth information before they begin crawling [24], the development of depth perception occurs typically after the sixth month [25].

Given the premature nature of depth perception, this work presents a mechanism of reaching primarily based on the progressive association between the hand position in iCub's gaze space V and the robotic arm's motor space M (please refer to Table I for a quick reference to symbols). The former is a two-dimensional space where the visual perception of the hand is located, whereas the latter is a four-dimensional structure for three shoulder and one elbow joints on iCub [2]. As part of the experimental machinery of the architecture, both gaze and motor spaces are further associated with the three-dimensional egocentric space K of the robot, where the x, y, and z-axes define the position of the hand with respect to robot's reference frame [26]. That is, the K space utilises the kinematics of the iCub to convert the joint coordinates from motor space into Cartesian coordinates and is used as an intermediary layer between the motor space and the gaze space.

Algorithm 1 Staged learning of reaching by hand regard

```
repeat
    $v$ and $v' \in V, m \in M$ and $k, h \in K$
    $vel \leftarrow \{\emptyset\}$
    $k_{x,y,z} \leftarrow getRandomCoordsInK()$
    $t_{x,y,z} \leftarrow handPositionInK()$
    while $dist(k_{x,y,z}, t_{x,y,z}) > 0.$ and $armNotStuck$ do
        $vel \leftarrow deriveNextVelocities()$
        $applyVelocities(vel)$                    ▷ Move robot's arm
        $t_{x,y,z} \leftarrow handPositionInK()$
    end while
    $h \leftarrow learnKfield(t_{x,y,z})$
    $m \leftarrow learnMfield(hand_{j0\rightarrow j3})$       ▷ Hand encoder

    $v' \leftarrow linkedVfield(h)$
    if $v'$ not $null$ then        ▷ Saccade to $v'$ if previously learned
        $saccadeVfield(v')$

    $saccadeRetina()$                 ▷ Refine head configuration
    $v \leftarrow learnVfield(head_{j0\rightarrow j5})$        ▷ Head encoder
    $associate(m, v, h)$
    $K_{pop} \leftarrow learnedKfields()$
until $K_{pop} = K_{stop}$                ▷ Empirically set to 200
```

When the robot fixates on a visual feature that belongs to its hand with both eyes, a field $v \in V$ is activated and associated with its equivalent motor field $m \in M$. For most objects, the visual features that constitute them may or may not fit within the same region in V, represented by v. In that case, v is seen as a part of the object. Subsequently, knowing the head configuration, i.e., the neck's pitch and yaw, as well as the eyes' tilt and version joint values, a field $k \in K$ is activated and linked to represent the position of the hand in the egocentric space. Note that version is regarded as the conjugate eye movement that accompanies saccades and smooth pursuit, in contrast to vergence, which causes the eyes to move in opposite directions. Algorithm 1 allows the population of these maps and the corresponding mappings. Moreover, each hand has its own reaching space that is learned separately. This is

[2] As previously noted, this decision depends purely on the body structure of the embodied agent.

necessary to avoid overlapping areas directly in front of the robot; a safety precaution taken to avoid collisions. Note that although K is acquired as part of the reaching machinery, the position of the hand throughout the experiments is found either by its visual or proprioceptive perceptions, a methodology which remains developmentally plausible.

As previously stated, infants spend a lot of their early life, observing the effect of their arms' motor babbling. They exhibit a pre-reaching movement repertoire where they stretch their arms to the extremes of the ranges of their joints towards objects but without being able to reach them [27], [28]. Soon, infants' nervous systems mature to allow the development of reflexive pre-reaching to visually elicited, cognition-driven reaching towards targets [29]. Similarly to vision, where the underdeveloped eyes are simulated, mimicking the transition of the arm movements in this work is achieved by considering a timeline of constraints, summarised in Table II. Notice that this table is an attempt to demonstrate a smooth developmental phase transition from reflexive movements to those elicited by vision. For instance, in the 2^{nd} stage of reaching development, it is anticipated that 60% of the arm movements are due to reflexes and only 10% have the potential to reach towards an observed target. Nevertheless, the success of this 10% of attempts is not guaranteed, as the reaching related maps are not fully populated, causing non-refined reaches. In turn, that also affects the system's ability to grasp, a reflection of the latter's dependency to the reaching movement [30].

TABLE II: Timeline of arm movements.

Stage	Arm movement (%)	Hand posture
1 (1-7 weeks)	R:100 E:0 V:0	open
2 (8-9)	R:60 E:30 V:10	open/closed/open
3 (10-11)	R:40 E:50 V:10	open/closed/open
4 (12-13)	R:20 E:60 V:20	open/closed/open
5 (14-15)	R:10 E:30 V:60	open/closed/open
6 (16-17)	R:0 E:0 V:100	open
7+ (18+)	R:0 E:0 V:100	open

R: reflex-only, E: elbow locked, V: visually elicited

Starting by randomly generated coordinates of the next target position to reach $k_{x,y,z}$ and by knowing the current hand position $h_{x,y,z}$ within the reach space K, the algorithm iteratively tries to minimise the Euclidean distance until no more movement is possible. At each iteration, the next commands to be sent to the motors derive from combined knowledge already acquired and stored as fields in the reaching space map as described in [8]. In brief, by subtracting the current $h_{x,y,z}$ from $k_{x,y,z}$ a vector T is derived which is then used to provide directional information to the target. All learned fields in K are then activated with respect to their distance to $h_{x,y,z}$. This process ensures that the information of fields closer to $h_{x,y,z}$ bears a more substantial weight than the distant ones. Utilising T to calculate the angle towards the target position, a weighted vector averaging is performed to estimate a new set of local velocity commands that move the hand closer to $k_{x,y,z}$. Note that initially when the maps are empty, the velocities are randomly generated as a result of motor babbling to allow the robotic arm to change position in the reaching space. Although this may lead to mistakes while reaching, appropriate fields are created and linked together for later use.

After several iterations and depending on the current population of K, the hand is expected to reach a position very close to the target $k_{x,y,z}$. Note that the less populated this space is, the less accurate reaching will be performed due to the weighted vector averaging. The resulting hand position is in turn used to learn (or activate an already learned) $h \in K$ that reflects the hand in the ego-centric space of the robot with respect to iCub's reference frame. Similarly, an $m \in M$ is learned to store the encoder values of the current arm configuration. However, what is necessary from a developmental point of view is to know the hand position in the gaze space, thus to learn a field $v \in V$ that encapsulates the position of the hand as perceived by the two eyes. A saccade to the hand by either reusing a previously learned $v' \in V$ that can bring the gaze to the vicinity of the hand, i.e., $saccadeVfield(v')$, or by performing a saccade in the retina as described in section II-D. After the head configuration changed and the eyes now fixate the hand, the algorithm is capable of associating m, v (or v') and h for the new proprioceptive, visual, and ego-centric information respectively. Finally, the number of fields learned in K, K_{stop}, is used as the stopping criterion of the hand-regard learning algorithm and is empirically set to 200 for each hand.

F. Learning hand postures while grasping

The ability to utilise both sensor and motor spaces of the upper limbs facilitates the learning of the physical body of a person in space even before birth [31]. The longitudinal development of infants' prehension has been thoroughly studied since the early 30s [32]. The results document the course of significant improvements infants undergo as they move from being equipped with a diversified motor reflex repertoire [33] to attempting visually-directed reaching and grasping techniques for some objects [22] and to developing sophisticated strategies that refine pre and post postures in achieving prospective grasp control [34]. In this work, employing grasping as a modality that enhances the understanding of objects is investigated. Tactile, as well as proprioceptive sensory information, are combined to exercise power gripping on objects. The grasping patterns, i.e., the hand posture once an object is securely grasped, are used to facilitate multi-modal object recognition. Algorithm 2 shows the power-grasping mechanism and the associated data capturing.

Power-grasping starts with an open hand and involves the thumb and all fingers. It is performed when any tactile equipped area on the hand receives a signal (reflexive response), and also when a visually elicited reach and grasp are required. At every iteration, the tactile module denoted as *"TactileLib"* reports on any haptic sensation being received on any of the five digits of the closing hand and updates the boolean set fts accordingly. This ensures that joints associated with touching fingers receive zero velocity for the next iteration. Subsequently, all non-touching fingers' joints receive a next velocity proportional to how close they are with respect to their maximum value. This technique ensures

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TCDS.2020.2965985, IEEE Transactions on Cognitive and Developmental Systems

Algorithm 2 Reflexive grasping

```
Require: A visual perception $v \in V$ that reflects the target object
$w, c \leftarrow 0$
$vel \leftarrow \{\emptyset\}$                          ▷ Next velocity set
$enc, enc' \leftarrow \{\emptyset\}$                   ▷ Motor vectors for 8 joints
$m \in M$
$n \leftarrow size(J)$
$fts \leftarrow \{$ false , false , false , false , false $\}$
repeat
    $fts \leftarrow receiveTactile()$
    $enc \leftarrow handEncoders()$
    for each hand joint $j \in J$ do
        if $fingerTouched(j, fts) =$ true then
            $vel_j \leftarrow 0$
            $n \leftarrow (n - 1)$
        else
            $diff_j = fabs(enc'_j - enc_j)$
            if $diff_j > jThresh$ then                ▷ Has moved
                $w \leftarrow fabs(enc_j - enc_j^{MAX})$
                if $w > 0$ then
                    $vel_j \leftarrow (w * .5)$
                else
                    $vel_j \leftarrow 0$
                    $n \leftarrow (n - 1)$
    $enc' \leftarrow enc$
    $sendVelocityCommands(vel)$
until $n = 0$
# gripper now closed
$enc \leftarrow handEncoders()$
$m \leftarrow learnMfield(enc)$
$c \leftarrow calculateGripClosure(enc)$
$associate(m, v, c)$
```

that the closer to the hand closed posture the fingers are, the slower movements are performed. Ultimately, when all fingers have stopped moving due to a blocking surface, the grasp is finalised and the algorithm records the encoders' values. The latter are used to learn a grasping configuration motor field $m \in M$ associated with the visual perception $v \in V$ that was previously learned by a fixation. Note that this time, M refers to the motor space of the hand. A more detailed study regarding the effectiveness of using proprioception and tactile sensory information to distinguish between grasped objects by iCub's hand is found in [35].

G. Building proto-objects

Casati describes proto-objects as operational objects that are distinct from the background and traceable throughout visual tasks [36]. They are representations of perceptions of objects in the scene that consist of salient features which share a consistency of motion. Therefore linking them offers an indexing system capable of referencing objects across the scene [37]. In this work, building such a representation is a process of employing the capabilities of the system described in sections II-C and II-D. Visual targets that result from camera image analysis and feature extraction are the building blocks of the process.

Proto-object generation makes use of maps to represent, and link together, fields in feature spaces. These are: (I) A two-dimensional motion map defined by the normalised area and the direction, with values ranging from 0–100 and 0–360 respectively. (II) A two-dimensional brightness map defined by the range of the Value element of pixels in the HSV model (0–255) and the normalised area of the region, ranging between 0–100. (III) A three-dimensional colour map defined by the values of the Hue, the Saturation and the normalised area, with ranges 0–180, 0–255 and 0–100, respectively. Note that the typical Hue range is scaled down, due to the use of OpenCV for image processing. Finally, (IV) a three-dimensional edges map defined by the values of the perimeter of the identified region, as well as its horizontal and vertical distances, with all ranges set to 0–500.

The radius of the fields in each feature map plays an important role in triggering the same field when targets with closely related properties are identified. The larger the field radius, the less accurate the system becomes in differentiating between features. On the contrary it becomes more tolerant to noise. The radii are set to 20, 25, 25 and 20 units, respectively. More details about the values used for feature representation in maps are found in [16].

Learning a proto-object involves the creation of a network of visual feature pairs that reflect what is currently observed within the scene. Features that share the characteristic of motion are assumed to have an association with the same animated object and thus can be paired [38]. It is worth noticing that for a single observation (e.g., feature extraction of an input image) on the retina, several instances of the same feature may exist. Consecutive observations are considered to ensure that only the salient features are associated in pairs whilst being aligned with the consistency in motion principle. The shortest distance d_{min} between all consecutive instances of a feature is calculated, and if $d_{min} < 40$ pixels it is said to be salient and can participate in pairs. Note that once linked, participating features and pairs constitute the abstractions of reusable information in the system and are stored in memory.

Linked pairs of salient feature instances are deemed as proto-objects (i.e., object representations) when a confidence level associated with a co-occurrence event is exceeded. Event heuristics are as follows:

- Appearance E^A is set to 1 when the two feature instances appear together. Otherwise set to 0.
- Movement E^M is set to 1 when their retina fields change at the same moment of time. Note that the direction of movement is not considered.
- Movement similarity E^S is set to 1 when the mean distance d_μ and standard deviation d_σ between two features is kept globally the same while $E^M = 1$. Note that monitoring is performed for a given amount of time and reliability in distances is measured by $|d - d_\mu| < 0.2 \times d_\sigma$, with d being the perceived distance between the features in the retina.
- Disconnection E^D is set to 1 when the mean distance d_μ is found to widely fluctuate.

Notice that the event heuristics above apply in the instances of features and pairs that are currently perceived in the retina and not on those that are learned.

The confidence value is used as an expression of strength for both the pair instance S^i (i.e., currently seen) and its pair abstraction S^p (i.e., saved in memory), written as:

$$\text{confidence} = S^p \times S^i \tag{1}$$

with S^i and S^p calculated by:

$$S^i = E^A \times w^A + E^M \times w^M + E^S \times w_i^S - E^D \times w_i^D \tag{2a}$$
$$S^p = E^S \times w_p^S - E^D \times w_p^D \tag{2b}$$

The weights w^A, w^M, w_i^S and w_i^D are empirically set to 0.01, 0.1, 0.9, and 1, respectively. Additionally, w_p^S and w_p^D are set to 0.01, 1, respectively. Note that w_p^S is significantly smaller that w_i^S in order to ensure that the learning rate is low for the pairs but high for their instances. The first time each pair is observed, S^p is set to 0, whereas S^i is set to 0 every time a new instance is created.

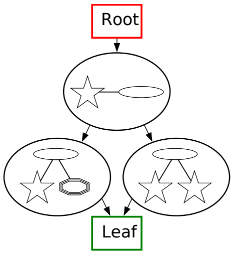


Fig. 3: An encapsulation graph. Stars, ovals and oval octagons represent colour, brightness and edges features respectively.

Thus far the system identifies salient points and uses event heuristics to express confidence for instances of pairs and features. As long as pair instances of a certain confidence (set to 0.5) have at least one common feature instance, they are used to formulate graphs, as seen in Figure 3. These are used to encapsulate the observed information and to facilitate future comparisons between any newly received and previously learned object perceptions. Comparing the two graphs renders the system capable of performing visual object recognition.

$$M^o = N^c / N^o \tag{3a}$$
$$M^m = N^c / N^m \tag{3b}$$

The M^o and M^m above capture the degree of similarity between the graphs. N^c, N^o and N^m are i) the number of common features between them, ii) the number of feature instances currently observed, and iii) the number of feature instances found in the previously learned observation, respectively. M^o is used to represent how close a feature graph in memory is to what is currently observed, and M^m to show how close a current observation is to what is stored in memory. Solving for M^o and M^m in equations 3a and 3b leads to some interesting observation phenomena; $M^m < 1$ and $M^o = 1$ implies that what is currently observed does not fully match the memory but the latter does match what is currently observed. In the contrary, $M^m = 1$ and $M^o < 1$ implies that what is currently observed is a supergraph of what is in memory. These two cases occur when an object is partially recognised. Consequently, $M^m = 1$ and $M^o = 1$ implies that there is a complete match, and $M^m < 1$ and $M^o < 1$ that there is no intersection between the graphs. The following measurement, designed to favour the perception of ambiguous recognitions, is proposed in order to provide a global recognition value of proto-objects:

$$M = 0.9 \times M^o + 0.1 \times M^m \tag{4}$$

Learning a new object is a result of capturing several visual perceptions that all represent some visual knowledge associated with it. Depending on the camera angle, object orientation, lighting conditions, etc., multiple graphs populate the memory. Those that best represent subsets of graphs associated with an object, thus encapsulating as much information of it as possible, are most valuable for recognition. In the context of this work, they are called proto-objects (denoted by PO), have a unique ID and a feature graph G^{PO} that best represents an object observation in memory. Note that many POs, each having an associated G^{PO}, can be recognised while the agent interacts with the object, or its position and orientation change. A more detailed description of the proto-object building mechanism is given in [39].

Although learning an object is a product of the co-occurrence of its feature instances in the retina, recognising a previously learned object should not depend on the object's movement. The mechanism of recognition should be capable of matching what is currently seen with the memory as if the object is moving. Either placed on a table or being partially hidden due to an obstacle, objects should trigger the same mechanism of co-occurring features, which is the prerequisite for understanding their associations. Evidence exists to demonstrate the importance of continuous small fixational eye movements in object perception, such as microsaccades [40]. In particular, researchers report that even with perfect retinal stabilisation, the human eye will be prone to fading of visual features due to an effect known as neural adaptation [41]. It is shown that suppressing small ocular movements leads to considerably larger changes to drifts and tremors of the eye, which in turn receives less visual stimulation [42]. Drawing inspiration from these studies and in order to recognise stationary objects, simulated movements of feature instances in the retina are produced. The system considers the result of these simulations as the currently observed information the equation 3. Hence, it can continuously perceive essential features for the generation of POs, even when fixating its gaze on an immobile object.

III. Multi-modal Object Learning and Recognition

Here, an infant-inspired multi-modal object learning and recognition process is explained. The parallel utilisation of the previously described components of the system facilitates learning behaviours and skills through observations and interactions with the objects within the environment. Note that the iCub's hand is considered as an object that the system can visually learn, and therefore recognise later. Like infants, the development of iCub as an embodied agent happens in stages, which differ in sensorimotor efficacies related to saccading between visual targets and reaching towards them.

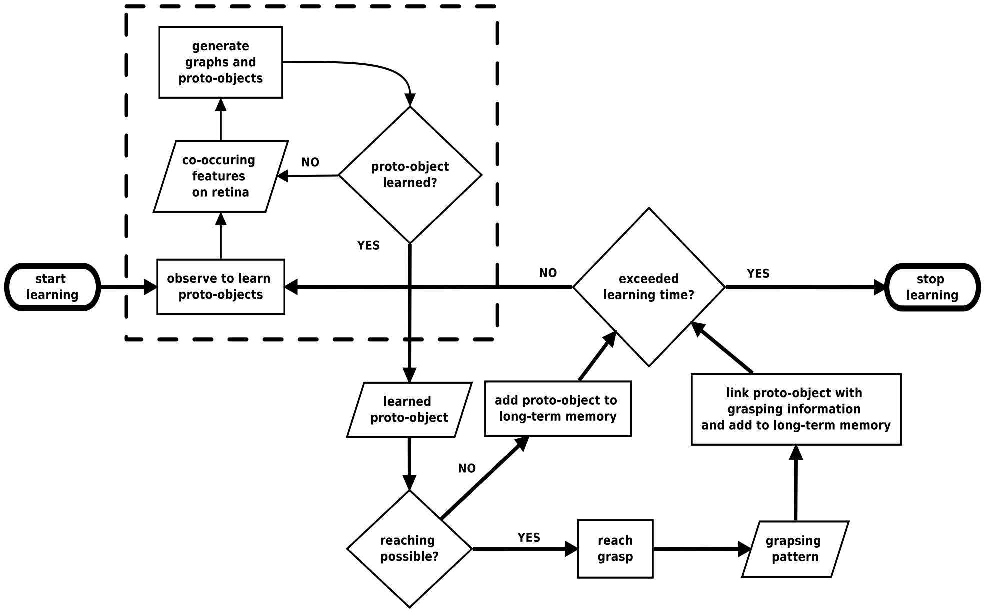

generate graphs and proto-objects
proto-object learned?
NO
co-occuring features on retina
YES
start learning
observe to learn proto-objects
NO
exceeded learning time?
YES
stop learning
learned proto-object
add proto-object to long-term memory
link proto-object with grasping information and add to long-term memory
NO
reaching possible?
YES
reach grasp
grapsing pattern

Fig. 4: The process of multi-modal learning of objects as a result of acquiring visual (i.e., proto-object) and grasp-related information about them. The process utilises the agent's vision as well as reaching and grasping modules as shown in Figure 1. Dashes frame an iterative process, during which movements of those features currently in the retina are simulated.

Appropriate maps and mappings for eye, neck and arm coordination are used at each developmental stage to reflect the improvement of skills as the agent matures. As described in section II-E, the more time is given to explore the reach space, the more it is populated with visual and proprioceptive observations of the hands. In order to mimic a transition to the arm movements, the timeline described in Table II is considered.

Both moving hands or objects are learned by the process illustrated in Figure 4. iCub is presented with a moving object to learn and ultimately recognise. The object can be either a toy or the robotic hand. In the case of the former, an experimenter constantly moves the object in the FOV, whereas for the latter the hand moves according to the timeline. Note that it may not always be in the FOV, making its learning and recognition a slow, challenging task.

While the object moves in front of iCub, graphs are learned to reflect the associated features that co-occur. The eyes then fixate the centroid of the features that constitute the object (and participate in the graph of the learned PO). If the feature is seen within reach, the arm and hand are engaged to reach and grasp the object. It is assumed that the system is intrinsically motivated to further interact with the object in order to acquire more information about it. The grasping pattern as defined by the proprioceptive information is associated with the currently learned or recognised PO, adding to the object knowledge. The interactions continue until the learning time threshold is reached. Ultimately, iCub is expected to have learned a number of POs, each one representing a different visual perception of an object, and a number of hand postures associated with it. That reflects familiarity of the system with grasping it and is limited to a threshold. It is assumed that after the threshold is reached per PO, the reaching and grasping action towards an object becomes habitual and not necessary.

The number of POs learned depends on the complexity of the observed object. Depending on the developmental stage of the system, which directly affects its ability to perceive the world through the cameras, the more co-occurring features the more POs are generated. Similarly, the more developed the eyes, the more complex in number of features the graphs are in memory.

After having learned an object, recognition can be performed. The experimenter places familiar objects in front of the robot one by one, and holds them still for some time. As previously mentioned, object recognition is a result of simulating movements of feature instances in the retina as a trigger to the matching mechanism. Now, the system is expected to generate a graph that matches what is currently observed as if it was moving, and to match it with some POs in memory. Note that the number of matching POs may differ for each object, and they are expected to have different confidence levels. Depending on the developmental stage, the difference between confidence levels may be noticeably small, leading to an ambiguous recognition. In this case, employing another modality to interact with the object is necessary.

The process of multi-modal recognition is shown in Figure 5. Given that there exist instances of feature pairs in the retina, the mechanism first simulates their movement to generate graphs for matching. When one or multiple matches are achieved the list of PO candidates is returned. Note that currently the system cannot dynamically switch between learning objects and recognising them, thus it is assumed that a PO memory already exists. At this point, each PO candidate consists of a list of features and their positions in the retina, as well as a confidence level. On the one hand, the retina

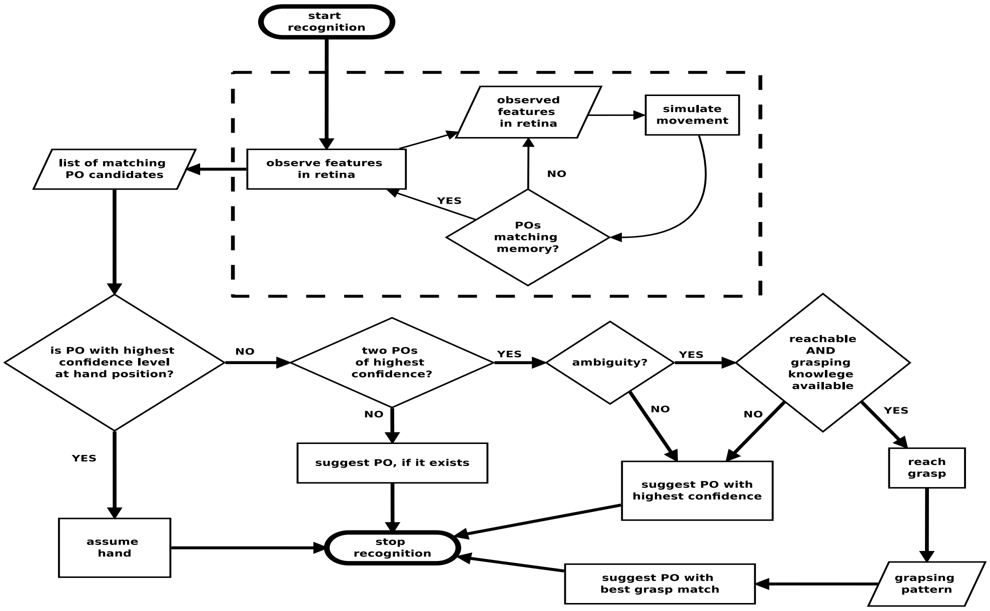

start recognition
observed features in retina
simulate movement
list of matching PO candidates
observe features in retina
NO
YES
POs matching memory?
is PO with highest confidence level at hand position?
NO
two POs of highest confidence?
YES
ambiguity?
YES
reachable AND grasping knowlege available
NO
NO
NO
YES
YES
suggest PO, if it exists
reach grasp
suggest PO with highest confidence
assume hand
stop recognition
suggest PO with best grasp match
grapsing pattern

Fig. 5: The process of multi-modal recognition of objects as a result of employing both visual and grasp-related information. Dashes frame an iterative process, during which movements of those features currently in the retina are simulated.

Algorithm 3 Hand posture matching

```
Require: $p_1$ and $p_2$ be the two ambiguous recognition candidates
Ensure: The return of either candidate $p_1$ or $p_2$
diff1, diff2 $\leftarrow \infty$
$g \leftarrow gripperInPerc()$            ▷ Hand posture after grasp in %
for each $p \in P$ do                     ▷ Set of POs previously learned
    if $p$ not hand then ▷ p is not associated only with the hand
        if $p == p_1$ then
            diff $\leftarrow |g - meanGrippersInPerc(p_1)|$
            if (diff < diff1) then
                diff1 $\leftarrow$ diff
                $cp_1 \leftarrow p_1$
        if $p == p_2$ then
            diff $\leftarrow |g - meanGrippersInPerc(p_2)|$
            if (diff < diff2) then
                diff2 $\leftarrow$ diff
                $cp_2 \leftarrow p_2$
return (diff1 < diff2) ? $cp_1$ : $cp_2$
```

positions are used to get the centroid of the whole object, important information for the understanding of depth, thus to know if it is within reach. On the other hand, the confidence levels dictate whether visually observing the object is enough to make a confident recognition.

The system utilises the proprioception of its arm joints and the hand regard associations it has learned to identify the hand position in the visual space. An assumption here is that if the position matches or is very close to the object's position, what is visually perceived is the hand. Note that the saccading module described in section II-D is responsible for translating the position of an object in the gaze space to the reaching space. Contrariwise, the process can either recognise the object by using only visual information or attempt to reach and grasp the object in order to resolve an ambiguity. The latter is attempted only if the object is found to be within the reachable space. Ambiguity occurs when there are more than one candidate POs and the ratio between the confidence level of the first two is higher or equal to an empirically pre-defined threshold, as shown in the conditional equation 5.

$$ambiguity(cPO_1, cPO_2) = \begin{cases} 1, & \text{if } (cPO_2/cPO_1) >= 0.8 \\ 0, & \text{otherwise} \end{cases} \tag{5}$$

where cPO_1 and cPO_2 are the confidence levels of the two POs respectively. Note that if there is only one PO in memory, then it is suggested for recognition. The stopping condition is also met when there is no memory of POs available.

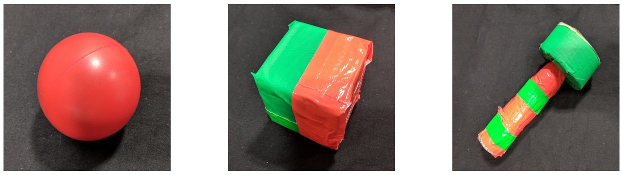
Grasping an object to acquire proprioceptive information is found to be a reliable mechanism to recognise objects of different sizes [35]. For this multi-modal process it is assumed that a successful match between hand posture patterns of what is currently grasped and what is in memory factors out the visual ambiguity and facilitates a confident recognition. Note that for every familiar object, iCub has collected several grasping patterns during learning. Thus, it is expected to have sufficient understanding of the object's shape to easily distinguish it from others. Algorithm 3 describes the recognition of objects based on hand posture matching.

IV. Experimental Methodology

The results of a longitudinal study whilst iCub experiences developmental changes to its abilities to visually perceive the world, saccade and reach towards objects are presented (Table III summarises the experimental settings). The robot is placed in a laboratory with typical lighting conditions. At every developmental stage, the system undergoes a learning phase during which iCub is given time to perform hand regard. Due to the simulated reflexes (described in section II-E), iCub builds proto-objects and associates them with the repositioning hand. Next, an experimenter presents three objects, one by one. Namely a red ball, a two coloured cube and a two coloured hammer, as shown in Figure 6. The experimenter constantly rotates and moves the presented object, making sure that it remains within the FOV of the agent. After both hand and objects are learned, the learning phase is over and the testing phase begins. Now, the experimenter places one object at a time in the iCub's FOV to be recognised. For each object 5 recognition trials are performed, each time placing the objects in random positions, far and close to the robotic hand.

TABLE III: Experimental settings.

Name	Value
learning time (secs)	120 (hand) 60 (objects)
ambiguity threshold	0.8
no. of grasps (habituation)	5 per PO
no. of recognitions	5 per objects
no. of developmental stages	10
objects used	hammer, cube, ball



(a) ball (b) cube (c) hammer

Fig. 6: Toys used in the experiment; different in shape, softness and colour.

The learning and testing phases are repeated for each developmental stage, with body maps being progressively updated to reflect the system's maturation. The following performance metrics are measured to evaluate the system:

- The percentage of time the hand is observed during hand regard. This is applicable only during hand regard as an indicator of how much time the system has generated POs while the hand repositions in the reaching space. This metric is negatively affected by; i) the amount of time the hand is physically placed within the FOV and ii) the ability of the eyes to perceive instances of features and pairs sufficiently for the creation of PO graphs.
- The number of reach requests and successful reaches during the learning phase of each stage. The earlier the stage the less requests will be successful due to the reflexes dominating the arm movements. However, in later stages iCub is expected to gain better control over its arms.
- The number of POs per object presented. The more visual features available, the larger the number of POs for each real object. The developing vision is expected to affect the number of POs being generated at each stage.
- Number of successful or failed recognition attempts and the associated reasons. This quantifies the contribution of using two modalities to recognise familiar objects.

V. Results and Discussion

iCub recognises its own hand after having learned a few POs while observing it moving (see Figure 7). The number of these new POs is not significantly large at the beginning as seen in Figure 9, due to the underdeveloped vision and the narrow FOV that causes the hand to frequently be outside the visual space. In stage 1, the number of POs associated with the hand allows iCub to recognise it for longer time during hand regard. The effect of utilising edge features as the vision develops is witnessed at stage 7 and 8, where the number of instance features currently observed in the retina significantly increases. As a result, more noise information is generated that has a negative effect on the generation of PO graphs to encapsulate the associations of salient and co-occurring features. At this developmental stage, the system suffers a decrease in the ability to recognise the hand, a characteristic clearly seen at the results of stage 7.

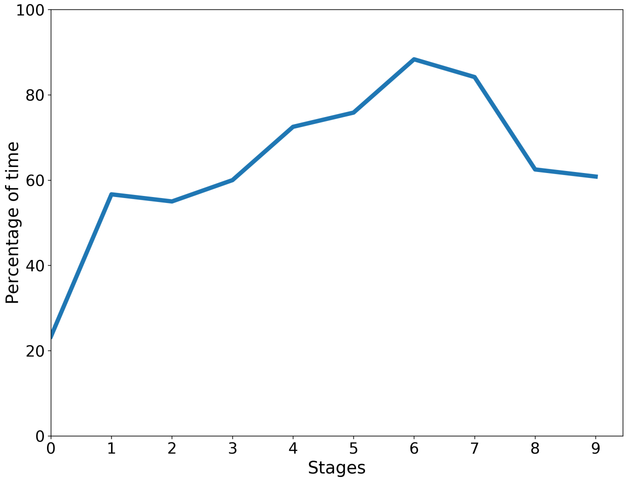


Fig. 7: The small number of POs generated as well as the narrow field of view due to the underdeveloped vision put a negative effect on the recognition of the hand.

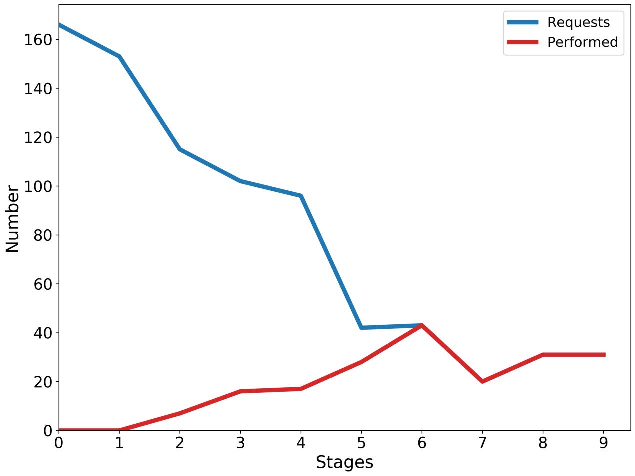


Fig. 8: The effect of reflexes coupled with the effect of reach and grasp habituation in physically interacting with objects during learning. After stage 6, every time iCub requests a reach towards an object it is found to be successful, rendering it able to recognise objects with confidence in a multi-modal fashion.

Every time a PO is generated or triggered by what is already in memory in order to reflect what is currently observed, iCub attempts a reach and grasp in order to acquire more information. It is assumed that the system obeys this intrinsic necessity to learn more until it has collected a sufficient number of grasping-related data. Furthermore, the success of each request does not depend only on the precision to reach towards a target, but also on whether the request is ignored due to reflexes taking over the arm modality. Figure 8 depicts the number of reach requests and successes during the longitudinal study. Unsurprisingly, the system requests more reaches to be performed with only a few of them being performed at the beginning due to the reflexes. The number of requests gradually drops as the system progressively learns enough hand postures for each learned PO related to an object. Notice that after stage 6, the reflexes are not affecting the ability of iCub to perform visually elicited reaches.

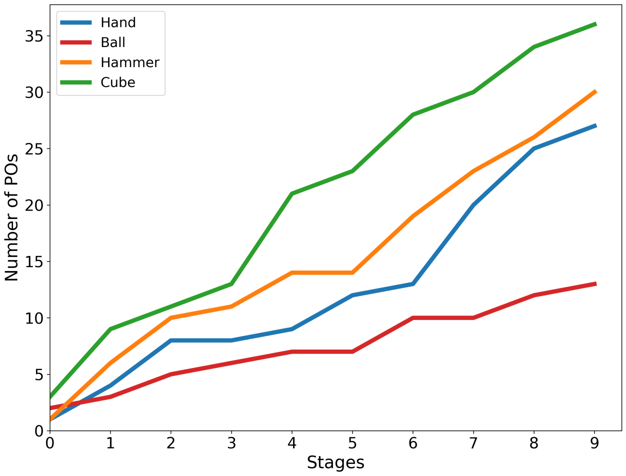


Fig. 9: The number new POs generated for each object per stage is found to increased as the vision capability of the system develops further.

The analysis of the results show that as more visual features are available due to the developing vision, the number of new POs generated for each object per stage also increases. The ability of the system to perceive edges is found to make significant contributions to the performance of learning and recognition. Figure 9 shows the number of POs generated per real object during the course of the development. Starting from the hand, it is seen that at stage 6 the number of POs representing different views of the moving hand increases. Considering the mechanical parts that constitute the segments of each digit on the iCub's hand and the amount of noise produced while it moves within the visual space, the system fails to recruit previously learned POs for what is currently observed. This justifies the drop in the percentage of time the hand is recognised in Figure 7. Note that the ball shows less complexity as it is, in fact, characterised by a few features regardless of the rotation dictated by the experimenter. From a developmental point of view, the poor vision in early stages stops the generation of multiple non-salient features that cannot be handled, rendering the system capable of performing ambiguous recognitions.

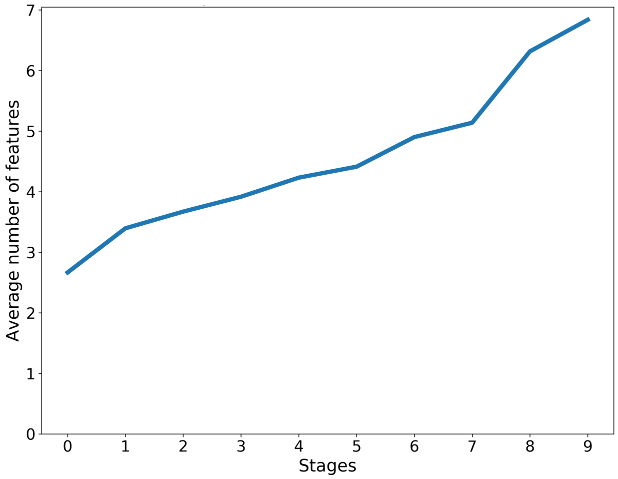


Fig. 10: The average number of features learned for all objects during the study. The complexity added due to the edge detection is depicted after stage 5.

Figure 10 depicts the average number of co-occurring features that are used to trigger or generate new POs during learning. Likewise, the complexity added due to the edges is depicted at stages 6–8, where new POs are encapsulating more complex representations of visual perceptions for both the hand and toy objects.

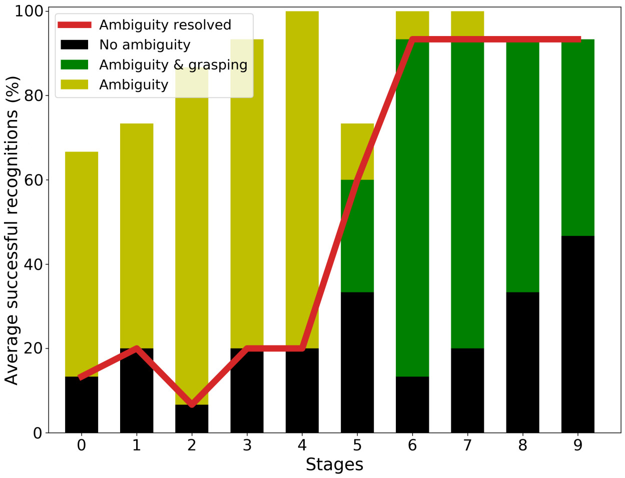


Fig. 11: The average successful recognitions for all the objects with stacked bars indicating how the system is able to successfully recognise objects at each developmental stage. With combined modalities, the system is capable of successfully recognising objects whilst increasing the level of certainty depicted by the red line.

Data collected during the testing phase of the experiment for each stage are summarised in Figures 11 and 12. Observing the average numbers of successful recognitions for all real objects in Figure 11, the coloured portions of each stacked bar indicate the reason of each success. That is, i) black, when the system is able to visually make a confident decision, ii) yellow, when ambiguity exists but an educated guess is successful, and iii) green, when the ambiguity is solved by reach and grasp.

The red line indicates the level of certainty in the system. It marks the performance when ambiguity was either not present or resolved by multi-modality. Although the system manages to recognise objects, it is found to lack certainty in making decisions during the first few stages. This is because most of the POs that are learned initially share a lot of similar features with each other. The impact of using both modalities in order to acquire further sensory information is seen after stage 5, where ambiguity is found to be resolved. Moreover, stage 5 appears to be a transition stage; a point in the iCub's development when incomplete grasping information exists that make the system prone to mistakes even after it has physically interacted with the object. The progress of the proto-object building is also shown at stage 6 onwards, when iCub is able to improve recognitions. The trend of the black portions is gradually increasing, meaning that more accurate (i.e., less similar to others) representative POs are generated and utilised. With combined modalities, the system can successfully recognise objects by minimising ambiguity.

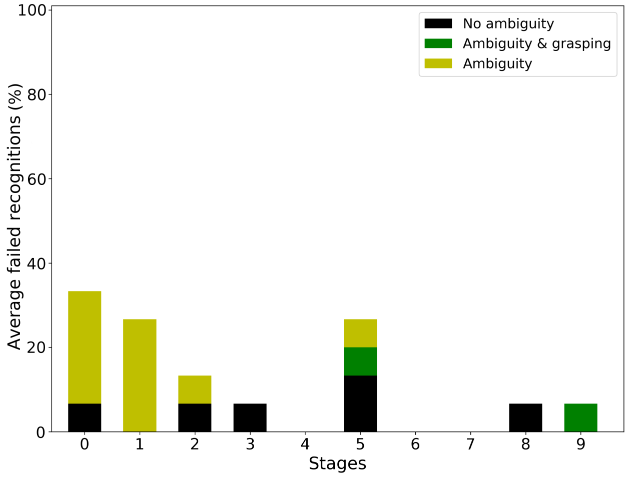


Fig. 12: The average failed recognitions during the course of the study. At stage 5, most of the failure is due to the underdeveloped vision (black portion). Interestingly, failures at this stage also occur due to the incomplete grasping information (green portion).

Figure 12 summarises the average unsuccessful recognitions and the reasons of failure at each developmental stage. It is seen that at stage 5, most of the failures are due to the underdeveloped vision (black portions). This is expected as the changes in vision after stage 4 have a stronger impact on the way iCub perceives the world. Edges are progressively identified and noise is picked up much easier by the vision module.

Failures also occur due to the incomplete grasping information the system has acquired for each PO. Although the average percentage of hand closure for each object make the objects distinguishable at later stages, incomplete information is found to lead to mistakes even after physical interactions. Table IV lists the hand posture results, with μ being the mean and σ being the standard deviation.

The results illustrate the impact of the developing sensor and motor control to the ability of the agent to recognise objects, including its own hand. In particular, the use of a developmentally plausible reflex system combined with physical as well as sight-related constraints explains the increased level of ambiguity in recognitions. It is shown that the initially poor vision acts as a natural defence against non-salient features that would cause ambiguities, a mechanism that is gradually suppressed as the system matures, making better use of other modalities. Consequently, the proposed system is capable of recognising objects in most of the cases, decreasing uncertainty as it matures.

VI. Conclusion and Future Work

This paper presents a system architecture that facilitates the longitudinal learning of a robotic embodied agent, following a similar approach to infant development. There are several key concepts in the architecture that form the basis of the proposed approach. These include sensorimotor learning for eye-head coordination, reaching and grasping based on hand regard, and a mechanism to build multi-modal object perceptions. Reported experiments demonstrate the ability of the system to acquire object-related knowledge by interacting with them, and its capacity to utilise this knowledge in order to perform successful recognitions.

TABLE IV: Average percentage of hand closure for each object.

Object	μ	σ
hammer	97 %	1.156
cube	55 %	0.267
ball	78 %	0.610

One important aspect highlighted in the results is the impact of ambiguity on the understanding of objects. Ambiguity is found not only in respect to the degree of resemblance between two or more different objects, but also between the competition of previously generated proto-objects that all partially match what is observed. In fact, a partial match is expected to be met more frequently than a complete one, due to the dynamic nature of the environment, e.g., partially hidden objects or slightly changed. This phenomenon is amplified as changes due to the developing vision occur between stages. Although able to recognise what is presented, the level of confidence does not increase until the reaching and grasping skills are mature enough to contribute. This is a product of scaffolding the learning of visual perceptions, whilst performing hand regard, and demonstrates the effect of multimodality in increasing recognition confidence. The results provide useful insights in the understanding of the refinement process of object knowledge found in infancy.

Our future work includes the extension of the system to dynamically switch between learning proto-objects during the recognition phase, and the integration with the high-level play generator module, labelled as *"Dev-PSchema"* in Figure 1. The combination of these two will facilitate further the investigation of the object learning processes and the discovery of their associated affordances as the agent interacts with the world, reporting any differences and impediments between the developmental stages.

Acknowledgements

This research is supported by the UK Engineering and Physical Sciences Research Council (EPSRC), grant No. EP/M013510/1.

References

[1] A. Gesell, “The tonic neck reflex in the human infant: Morphogenetic and clinical significance,” *The Journal of Pediatrics*, vol. 13, no. 4, pp. 455–464, 1938.

[2] J. Law, P. Shaw, K. Earland, M. Sheldon, and M. H. Lee, “A psychology based approach for longitudinal development in cognitive robotics,” *Frontiers in neurorobotics*, vol. 8, p. 1, 2014.

[3] J. Sinapov, T. Bergquist, C. Schenck, U. Ohiri, S. Griffith, and A. Stoytchev, “Interactive object recognition using proprioceptive and auditory feedback,” *The International Journal of Robotics Research*, vol. 30, no. 10, pp. 1250–1262, 2011.

[4] G. Metta, L. Natale, F. Nori, G. Sandini, D. Vernon, L. Fadiga, C. Von Hofsten, K. Rosander, M. Lopes, J. Santos-Victor, and A. Bernardino, “The iCub humanoid robot: An open-systems platform for research in cognitive development,” *Neural Networks*, vol. 23, no. 8-9, pp. 1125–1134, 2010.

[5] M. H. Lee, Q. Meng, and F. Chao, “Staged competence learning in developmental robotics,” *Adaptive Behavior*, vol. 15, no. 3, pp. 241–255, 2007.

[6] J. Law, M. Lee, M. Hülse, and A. Tomassetti, “The infant development timeline and its application to robot shaping,” *Adaptive Behavior*, vol. 19, no. 5, pp. 335–358, 2011.

[7] V. Braitenberg and A. Schüz, *Anatomy of the cortex: statistics and geometry*. Springer Science & Business Media, 2013, vol. 18.

[8] K. Earland, M. Lee, P. Shaw, and J. Law, “Overlapping structures in sensory-motor mappings,” *PloS one*, vol. 9, no. 1, p. e84240, 2014.

[9] D. Maurer, T. L. Lewis, H. P. Brent, and A. V. Levin, “Rapid improvement in the acuity of infants after visual input,” *Science*, vol. 286, no. 5437, pp. 108–110, 1999.

[10] D. Maurer and T. L. Lewis, “Visual acuity: the role of visual input in inducing postnatal change,” *Clinical Neuroscience Research*, vol. 1, no. 4, pp. 239–247, 2001.

[11] A. M. Brown, D. T. Lindsey, E. M. McSweeney, and M. M. Walters, “Infant luminance and chromatic contrast sensitivity: optokinetic nystagmus data on 3-month-olds,” *Vision Research*, vol. 35, no. 22, pp. 3145–3160, 1995.

[12] D. Y. Teller and M. H. Bornstein, “Infant color vision and color perception,” *Handbook of infant perception*, vol. 1, pp. 185–236, 1987.

[13] M. S. Banks and E. Shannon, “Spatial and chromatic visual efficiency in human neonates,” *Visual perception and cognition in infancy*, p. 146, 1993.

[14] S. P. Johnson, “How infants learn about the visual world,” *Cognitive Science*, vol. 34, no. 7, pp. 1158–1184, 2010.

[15] J. Canny, “A computational approach to edge detection,” *IEEE Transactions on pattern analysis and machine intelligence*, no. 6, pp. 679–698, 1986.

[16] D. Lewkowicz, A. Giagkos, P. Shaw, S. Kumar, M. Lee, and Q. Shen, “Towards learning strategies and exploration patterns for feature perception,” in *Development and Learning and Epigenetic Robotics (ICDL-EpiRob), 2016 Joint IEEE International Conference on*. IEEE, 2016, pp. 278–283.

[17] A. Giagkos, D. Lewkowicz, P. Shaw, S. Kumar, M. Lee, and Q. Shen, “Perception of localized features during robotic sensorimotor development,” *IEEE Transactions on Cognitive and Developmental Systems*, vol. 9, no. 2, pp. 127–140, 2017.

[18] J. Law, P. Shaw, and M. Lee, “A biologically constrained architecture for developmental learning of eye–head gaze control on a humanoid robot,” *Autonomous Robots*, vol. 35, no. 1, pp. 77–92, 2013.

[19] G. Butterworth and B. Hopkins, “Hand-mouth coordination in the newborn baby,” *British Journal of Developmental Psychology*, vol. 6, no. 4, pp. 303–314, 1988.

[20] P. Rochat, E. M. Blass, and L. B. Hoffmeyer, “Oropharyngeal control of hand-mouth coordination in newborn infants.” *Developmental Psychology*, vol. 24, no. 4, p. 459, 1988.

[21] P. Rochat, “The emergence of self-awareness as co-awareness in early child development,” *Advances in Consciousness Research*, vol. 59, pp. 1–20, 2004.

[22] B. L. White, P. Castle, and R. Held, “Observations on the development of visually-directed reaching,” *Child development*, pp. 349–364, 1964.

[23] A. Yonas, “Infants’ responses to optical information for collision: Psychobiological perspectives: The visual system,” in *Development of perception: Psychobiological perspectives: The visual system*. Academic Press, 1981.

[24] K. E. Adolph, “Specificity of learning: Why infants fall over a veritable cliff,” *Psychological Science*, vol. 11, no. 4, pp. 290–295, 2000.

[25] E. J. Gibson and R. D. Walk, “The “visual cliff”,” *Scientific American*, vol. 202, no. 4, pp. 64–71, 1960.

[26] U. Pattacini, F. Nori, L. Natale, G. Metta, and G. Sandini, “An experimental evaluation of a novel minimum-jerk cartesian controller for humanoid robots,” in *Intelligent Robots and Systems (IROS), 2010 IEEE/RSJ International Conference on*. IEEE, 2010, pp. 1668–1674.

[27] C. von Hofsten, “Developmental changes in the organization of prereaching movements.” *Developmental Psychology*, vol. 20, no. 3, p. 378, 1984.

[28] A. Bhat, H. Lee, and J. Galloway, “Toy-oriented changes in early arm movements II–Joint kinematics,” *Infant Behavior and Development*, vol. 30, no. 2, pp. 307–324, 2007.

[29] T. Homma, “Hand recognition obtained by simulation of hand regard,” *Frontiers in Psychology*, vol. 9, 2018.

[30] P. Zech, E. Renaudo, S. Haller, X. Zhang, and J. Piater, “Action representations in robotics: A taxonomy and systematic classification,” *The International Journal of Robotics Research*, vol. 38, no. 5, pp. 518–562, 2019.

[31] M. Hoffmann, “The role of self-touch experience in the formation of the self,” *arXiv preprint arXiv:1712.07843*, 2017.

[32] H. M. Halverson, “Studies of the grasping responses of early infancy: I,” *The Pedagogical Seminary and Journal of Genetic Psychology*, vol. 51, no. 2, pp. 371–392, 1937. [Online]. Available: https://doi.org/10.1080/08856559.1937.10532507

[33] J. Schott and M. Rossor, “The grasp and other primitive reflexes,” *Journal of Neurology, Neurosurgery & Psychiatry*, vol. 74, no. 5, pp. 558–560, 2003.

[34] D. C. Witherington, “The development of prospective grasping control between 5 and 7 months: A longitudinal study,” *Infancy*, vol. 7, no. 2, pp. 143–161, 2005.

[35] A. Giagkos, Raphaël, P. Shaw, M. Lee, and Q. Shen, “Assessing Humanoid Multimodal Grasping Towards Object Recognition,” in *2nd Robot Manipulation Workshop*, Imperial University, London, July 2017.

[36] R. Casati, “Object perception,” *Oxford handbook of philosophy of perception*, pp. 393–404, 2015.

[37] A. Clark, “Feature-placing and proto-objects,” *Philosophical Psychology*, vol. 17, no. 4, pp. 443–469, 2004.

[38] J. Fiser and R. N. Aslin, “Statistical learning of new visual feature combinations by infants,” *Proceedings of the National Academy of Sciences of the United States of America*, vol. 99, no. 24, pp. 15 822–15 826, 2002.

[39] R. Braud, A. Giagkos, P. Shaw, M. Lee, and Q. Shen, “Building Representations of Proto-Objects with Exploration of the Effect on Fixation Times,” in *7th International Conference on Development and Learning and on Epigenetic Robotics (ICDL-EPIROB 2017)*. Lisbon, Portugal: IEEE, September 2017.

[40] S. Martinez-Conde, J. Otero-Millan, and S. L. Macknik, “The impact of microsaccades on vision: towards a unified theory of saccadic function,” *Nature Reviews Neuroscience*, vol. 14, no. 2, p. 83, 2013.

[41] S. Martinez-Conde, S. L. Macknik, X. G. Troncoso, and T. A. Dyar, “Microsaccades counteract visual fading during fixation,” *Neuron*, vol. 49, no. 2, pp. 297–305, 2006.

[42] M. Clowes, “A note on colour discrimination under conditions of retinal image constraint,” *Optica Acta: International Journal of Optics*, vol. 9, no. 1, pp. 65–68, 1962.

Raphaël Braud received his MSc in Intelligent Systems and Robotics (2012) from the University of Cergy-Pontoise, France. He holds a Ph.D. in Developmental Robotics (2017) on the modelling of cognitive mechanisms for sensorimotor control and tool-use from the same institution. Dr Braud has recently finished working as a Post Doctoral Research Associate, conducting research on developmental learning for humanoid robots through interactions with objects and tools as part of the MoDeL project at Aberystwyth University. He is now working as a Research Engineer in Machine Learning at the French Institute of Technologie SystemX, France.

Alexandros Giagkos received the degrees of B.Sc. in Computer Science, M.Sc. in Internet and Distributed Systems and Ph.D. in Computer Science from Aberystwyth University. He was a Lecturer at Aberystwyth University and recently moved to Aston University, Birmingham. His main regions of research interest are developmental, evolutionary and swarm robotics.

Patricia Shaw received her B.Sc. in Artificial Intelligence (2005) and Ph.D. in Computer Science (2010) from the University of Durham. She has recently finished working as a Post Doctoral Research Associate, researching developmental robotics as part of the European Framework 7 IM-CLeVeR project, and is now a Lecturer in the Intelligent Robotics Group at Aberystwyth University. Her research interests include biologically and psychologically inspired architectures for developmental learning in robotic systems.

Mark Lee Prof. Lee received the degrees of B.Sc. (1967) and M.Sc. (1969) in Electrical Engineering from the University of Wales, Swansea, and Ph.D. (1980) in Psychology from Nottingham University. He is emeritus Professor of Intelligent Systems in the Department of Computer Science at Aberystwyth University, Wales, UK. His main research interests are in Developmental Robotics, particularly in relation to early infant psychology. He was Principal Investigator on four recent EPSRC and EC funded research projects on robotic sensory-motor learning, adaptation and development. He is a Fellow of the Learned Society of Wales.

Qiang Shen received the Ph.D. in Computing and Electrical Engineering (1990) from Heriot-Watt University, Edinburgh, U.K., and the D.Sc. in Computational Intelligence (2013) from Aberystwyth University, Aberystwyth, U.K. He holds the Established Chair in Computer Science and is the Pro Vice-Chancellor: Faculty of Business and Physical Sciences, Aberystwyth University. His research interests include computational intelligence and its application in robotics. He has authored two research monographs and over 390 peer-reviewed papers, including one receiving an Outstanding Transactions Paper Award from the IEEE.