Developmental robotics architecture for active vision and reaching

Martin Hülse, Sebastian McBride, Mark Lee
Department of Computer Science, Aberystwyth University, SY23 3DB, UK
Email: {msh,sdm,mhl}@aber.ac.uk

***Abstract*—A robotic architecture is presented which is inspired by the process of developmental learning of human infants in their twelve months after birth. The architecture integrates active vision and simple object manipulation (reaching and grasping). Robotic experiments demonstrate how visual and non-visual features determine visual attention and reaching. However, more important and main objective of this paper is the organisation of the architecture with respect to developmental learning of firstly, the behavioural competence of hand-eye coordination and secondly, the cognitive competence of multimodal visual attention. These topics of staged competence learning of behavioural and cognitive skills are discussed. In this discussion we also outline the value of this architecture as reference model for the investigation of mechanisms of staged competence learning in humanoid robots or even biological systems.**

I. Introduction

Creating humanoid robot systems that share and master the varieties of our human environment in order to assist and accompany us human beings in our daily work and private life could be referred to as the "Holy Grail" of robotics. In the field of Developmental Robotics (1) it is assumed that this ultimate goal is achieved best if such a robot system is not programmed for specific and fixed tasks, but if it is programmed to develop and learn new behavioral and cognitive competences and skills autonomously. Following further the argumentation of embodied cognition and embodied artificial intelligence (2), namely that intelligence is based on the physical interaction with the world and that cognitive abilities are shaped by the way a system is interacting with its environment, then human intelligence is heavily determined by our human bodies and the way we develop from birth. Under normal circumstances a human infant develops in characteristic behavioral and cognitive stages. It can be assumed that these developmental stages as the outcome of evolutionary selection refer to a developmental process that has evolutionary advantage for the survival of the human being in this world. Thus, the developmental stages point to strategies that are best suited to create control mechanisms mastering robustly the complex coordination of human-like sensorimotor systems. Thus, when faced with the challenge of engineering control procedures for humanoid robots it seems straight forwards to try to mimic the essential characteristics of human development because humanoid robots are intended to reproduce the essential sensorimotor qualities, or the embodiment, of humans. Therefore we argue, the developmental stages of human beings might be very helpful to find robust strategies for the autonomous development of "intelligent" or at least fairly useful humanoids.

Our research in Developmental Robotics is focused on the first twelve months of a human infant and therefore, the behavioral competences we try to reproduce on humanoid or anthropomorphic robot systems are active vision, visual attention, hand-eye coordination and simple object manipulation. We believe these are the basic building blocks that ground the multimodal object representations that later provide the computational substrate for cumulative and active learning, social learning as well as abstract knowledge representation and reasoning. Furthermore, we are interested in providing a *developmental framework* where these competences are learned autonomously following specific stages of human infants. This is in contrast to approaches, we call them here *non-developmental frameworks*, where single tasks or competences are explicitly programmed and calibrated in isolation. When this is done, these isolated modules are put together into one system by an engineer. In a recent study we have shown that developmental frameworks can compensate a crucial effect that is related to the integration of several sensorimotor competences (3). If a non-developmental framework is applied to learn different sensorimotor competences separately, but which have to be combined and integrated for a global task then there is an *accumulation of uncertainty*. This causes high uncertainty in the global coordination task although the single sensorimotor modalities have low error. This is because the global error is the result of the accumulation of the uncertainties all the single modalities inherently carry. A developmental framework can compensate the accumulation of uncertainty. Thus, the global error rate is low although single sensorimotor modalities might show large error rates. In conclusion, we can say that developmental frameworks lead to better sensorimotor coordination for robot systems that have to integrate several sensorimotor competences. Therefore, we argue that architectures for advanced robot systems having several sensorimotor modalities and high degrees of freedom need to apply developmental learning. From a purely engineering point of view this is not just for its own sake of applying learning mechanisms but more important because developmental learning is able to compensate the accumulation of uncertainty.

The objective of this paper is the introduction of a robotics architecture for developmental learning that enables humanoid and anthropomorphic robot systems to learn the competence of visual search and simple object manipulation. The architecture

978-1-61284-990-4/11/$26.00 ©2011 Crown

is the result from former robotic experiments on visual attention, eye-saccades learning and hand-eye coordination (4; 5). At this stage of research we can present this architecture as a validation of the framework only, meaning no developmental learning is directly involved. The learning of sensorimotor competences has already taken place and do not undergo any adaptation process anymore. However, the validation of the architecture fully working in a robotic setup shows already elements essential for a developmental learning framework, which we will discuss in detail later. On the other hand, we believe that the current architecture is worth presenting since it can be used as reference architecture for other research activities where other strategies for developmental learning are investigated.

In the following we present our computational architecture for active vision and object manipulation. We further present experiments where multimodal visual attention is demonstrated. This is followed by a discussion of the essential features that make this architecture suitable for the study of developmental learning.

II. Architecture

A. Robotic Setup

The active vision system is part of an anthropomorphic robotics system, where active vision is combined with a robotic manipulator. The active vision system consists of two cameras (both provide RGB 1032x778 image data) mounted on a motorised pan-tilt-verge unit. Here, only two DOF, verge and tilt of the left camera, are used. The motors are controlled by determining their absolute target position p, or the change of the current position Δp, given in radians (*rad*).

The active camera is faced towards the table the manipulator is mounted on. The objects on the table are coloured and they can be relocated by the manipulator.

1) Separation of visual "where" and "what" data: The image processing of the original RGB camera image data was set up to produce a basic representation of visual input data as seen within the human eye. This was simulated by dividing the original camera RGB data into two data streams; the first fed high resolution image data from a small localised region within the centre of the image (simulating the fovea, the center of the human retina), the second low resolution stream represented visual information outside of this region (perifoveal). The low resolution data were RGB filtered with the individual filter outputs linearly combined and normalised to generate a retinotopic map with each pixel in this map having the range $[0.0, 1.0]$ indicating the intensity of colour components red, blue and green. This is similar to the saliency maps (6).

Feature filters were applied to the high resolution image data only, in order to extract the exact intensity of each individual colour component and information about shape or texture. These data are summarised in form of a *feature vector* v_v.

Thus, the original RGB image data was transformed into two data streams: one delivering a low resolution retinotopic map using RGB filtering and the other a high-resolution-based feature vector v_v. The low resolution retinotopic map simulates the "where"-stream, while the feature vector derived from the high resolution data at the image centre simulates the "what"-stream. These two streams of visual data are applied to simulate the dorsal ("where") and ventral ("what") pathways of visual processing in the human brain.

2) Object fixations: Since visual features can only be detected from the image center, the camera must fixate the object in order to generate the visual feature vector v_v for this object. Fixation can be achieved by saccadic camera movements which bring a selected image region into the image center, as previously described (4). In brief, a peripheral stimulus located at (X,Y)-coordinate within the retinotopic map is linked, through a learning process, to specific relative motor movements Δp. The relative motor movements thus brings the stimulus to the image centre where the feature vector v_v can then be derived. The eye saccade is said to be successful if after the execution the image centre of the saliency map contains non-zero entries.

The linking between (X,Y)-coordinates and relative motor movements Δp is represented by a sensorimotor mapping which is learned previously. We call it *eye-saccade mapping*. The eye-saccade mapping is an essential prerequisite for the transformation of visual stimuli into the egocentric reference frame.

It is important to notice that each object fixation is fully determined by the absolute motor positions the active vision system ends after executing a saccade. Thus, each object can be associated by a unique motor absolute configuration p. The space defined by the absolute motor positions of verge and tilt motor is what we call the *gaze space*. As described later in detail the gaze space provides the egocentrics reference frame where integration of "what" and "where" data takes places. The usage of gaze space is also motivated by former experiments where we have shown that a robotic system can successfully learn hand eye-coordination by using the gaze space as absolute reference frame (5; 7).

B. Robotic architecture for active vision and object reaching and grasping

Figure 1 indicates the data flow between the three main computational domains of our robotic architecture: retinotopic, gaze and feature space. The latter, the feature space, includes reach, tactile and colour feature space.

As described above the image data delivered from the camera are represented in a retinotopic reference frame where we derive low resolution location data ("where") as well as feature data ("what") but from the image centre.

Feature data are represented in an abstract feature space containing no location data, while location data from the retinotopic reference frame are transformed into the gaze space which provides an absolute, our egocentric, reference frame. Notice, saccade generation is computed in the gaze space not in the retinotopic reference frame. This allows, as only one example, saccades towards objects which are not in the current visual field.

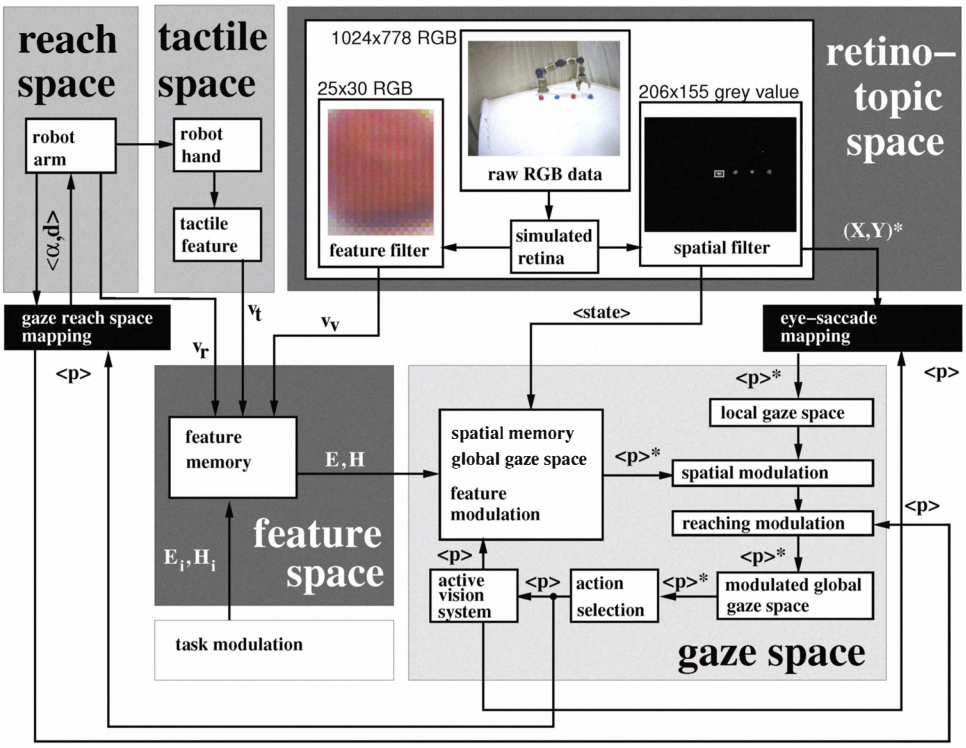

Fig. 1. Computational architecture for active vision and object manipulation.

1) Transformation from retinotopic to egocentric coordinates: The transformation from visual stimuli in the retinotopic reference frame $(X,Y)^*$ to gaze coordinates $< p >^*$ is based on the eye-saccade mapping by deriving the relative verge and tilt motor movements Δp for each stimuli (X,Y) and adding onto the current absolute motor positions of the active vision system p (5). The gaze space provides the domain for action selection of saccadic camera movements whereby the most salient stimulus p would drive the camera motors into position p.

Once the camera has successfully saccaded the selected stimulus at p then this gaze coordinate is stored in the *spatial memory*. Every entry in the spatial memory has an age value a measured in seconds. If this age value is larger M then the entry is removed from the memory.

2) Spatial modulation and inhibition of return: The spatial memory is used to modulate the current visual input coming from the spatial filter. Since both are represented in the absolute gaze space, stimuli in the spatial memory inhibit the stimuli coming from spatial filter if they have the same coordinates / location p in gaze space. Consequently, a stimulus in the current visual input that was already fixated is inhibited. Thus, a selection for this stimulus for the next eye saccade becomes more unlikely. The inhibition is stronger the smaller the corresponding age value a is. We call this process *spatial modulation*. Spatial modulation generates an inhibition of return mechanism (IOR) allowing the system to fixated every object in the scene without repeatedly saccading to the very same object while ignoring other objects.

3) Feature modulation: On top of spatial modulation there is feature modulation which is formally written as:

$$f(p) \quad = \quad s - \left(1 - \frac{a}{M}\right) \cdot \left(\frac{H+1}{E+1}\right), \tag{1}$$

where $f(p)$ is the activation value of p before the action selection process takes place. The value s $(0 \leq s \leq 1)$ is the original saliency value provided by the spatial filter. The values a and M $(a, M > 0)$ determine spatial modulation, where a is the age value of p in the spatial memory and M is the maximal age before p is removed from the spatial memory. Both values are given in seconds. Furthermore, $H, E \in \mathbb{R}$ and $0 \leq H, E$ determine the excitation (E) and inhibition (H) level. The E and H values are derived from the feature vector in the feature memory the stimuli p is associated with. Notice, if $E = H$ then we have spatial modulation only.

4) Feature association, binding "where" and "what" data: Feature association takes place during object fixation and reaching and grasping. Once the object is fixated, the current input coming from the feature filter v_v is stored in the feature memory. Furthermore, once the object is grasped by the robotic manipulator the corresponding tactile feature v_t and reach v_r coordinates can be stored in the feature memory as well. All of these feature values are link to the p coordinates just stored in the spatial memory since this entries represents the location in gaze space of the fixated and grasped object. As we have already mentioned the feature vectors associated or linked with the stimulus p in the spatial memory determine its E and H value.

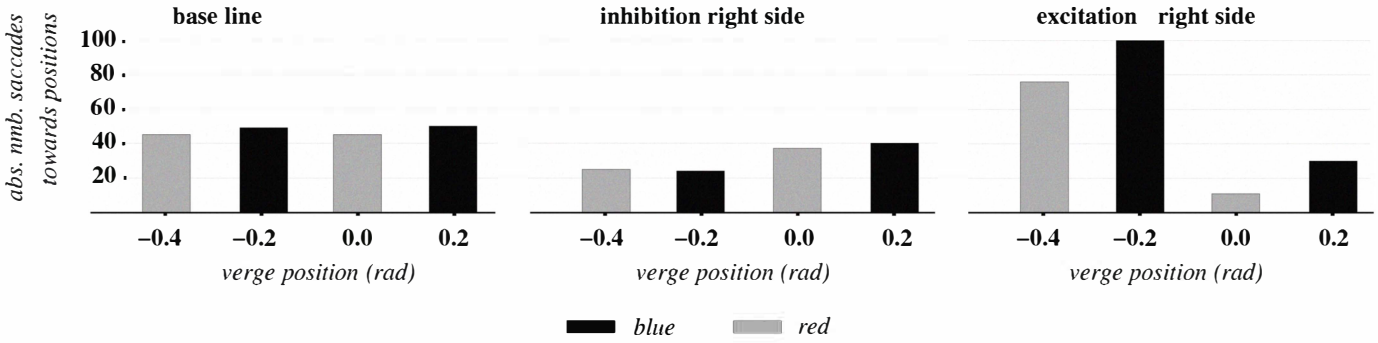

Fig. 2. Total number of saccades towards red and blue objects while spatial habituation takes place, recording time 600 seconds, $M = 20$. **Left:** spatial modulation only. **Middle:** Inhibition of the right side in reach space. **Right:** Excitation of the right side in reach space.

5) Task modulation: Feature modulation is determined by the E and H values which are again determined by the feature vectors p is linked with. In the case of more complex objects and feature vectors there are many possible strategies to derive the inhibition and excitation levels out of the individual feature values. For the sake of simplicity we provide experiments for disjunct colour, tactile and reach features classes only. However, this is sufficient to demonstrate the way low resolution visual input (spatial information) can be modulated by high resolution visual and non-visual features. The assigned E and H values for each feature class that determine the task relevance of a specific object. This assignment is what we call *task modulation* in our architecture.

III. Experiments

In the following we present experiments demonstrating two types of multimodal feature modulation. In all the experiments the capacity M of the spatial memory is 20 seconds.

A. Reach space feature association

Here, the system behaviour is measured in terms of fixation patterns. Over a period of 600 seconds the number of saccades is recorded. In addition, for each saccade we recorded the p-value and the received features class. Hence, for each saccade we know which object the system has actually fixated and the feature classes perceived.

In this scenario we have used feature associations with the reach space to bias the vision system towards the "right side" (called feature class RG) of the robot manipulator while being neutral $E = H$ for its left side (feature class LE). In this scenario on both sides (LE and RG) of the robot arm one red and one blue object is placed. Thus, in total four objects. Since reach location can directly associated with gaze space location, no reaching and grasping action is executed, and therefore the scenario is static. Excitation and inhibition values for colour were the same and does not need to be considered here further. However, we altered the relation between the E and H values for the reach space feature classes, namely the robots "left"(LE) and "right"(RG) side. Thus, the final excitation and inhibition values are calculated as follows: $E = E_{LE} + E_{RG}$ and $H = H_{LE} + H_{RG}$, where final excitation value E is the sum of excitation values assigned to the robots left and the right side; and vice versa for the final inhibition value H.

We present three runs. The first provides the base line, where there is no bias towards reach feature class LE or RG. This is formally written as: $E_{LE} = E_{RG} = H_{LE} = H_{RG} = 0$. In the second run, the system was biased towards the left side of the robot arm by inhibiting the right side only: $H_{RG} = 9$ while all the other values are zero. Therefore, we have $H = 9$ and $E = 0$ for all the stimuli p in the spatial memory which are associated with RG while all p associated with LE have $H = E = 0$.

Finally, in the third run we biased the system towards the robot's arm right side by excitation of the corresponding feature class RG: $E_{RG} = 9$ while all the others values are zero and therefore, we have $H = 0$ and $E = 9$; and again all stimuli p in the spatial memory which are associated with LE have $H = E = 0$.

The resulting fixation patterns are presented in Fig. 2. These diagrams show the number of saccades towards the four individual objects. The spacial position is indicated by gaze space coordinates. Absolute motor positions of the verge left motor larger $\approx 0.1 rad$ represent the left side of the robot arm (LF); while verge positions less or equal 0.1 represent the right side of the arm (RG). When inhibition or excitation of the right side takes place then the number of saccades towards the objects on the left and right side differ significantly. Since blue and red objects are on both sides it is obviously the spatial association that causes these differences. Hence, difference in the total number of object fixations between the manipulators left and right side can only be generated by non-visual spatial reach feature associations, not by the visual features. Hence, the position in reach space determines the fixation of objects, or in other words, the robot's "visual attention window".

B. Indirect tactile feature association

Cross-modal feature modulation is here demonstrated in a scenario where we have two red and two blue balls. The blue ones are soft and the red ones hard. Hence, while grasping them they can easily be classified by the system according to the two tactile features classes SO (hard) and HA (soft). The excitation values for the tactile feature classes are set as follows: $E_{SO} = H_{HA} = 0$ and $E_{HA} = H_{SO} = 9$. which expresses

a preference towards hard objects while soft should be avoided. All other excitation and inhibition values are zero.

The task of the system is to fixate an object, to pick it up and put it back on the table at a new position. Due to the *reaching modulation* the active vision keeps the object fixated while reaching and grasping are executed. After the object is placed back there is a time period of 35 seconds where the system is not allowed to trigger a reach action. After this period of time reaching and grasping actions are triggered as soon as an object is fixated. It is by chance which object is picked up since objects are repeatedly fixated. Thus, a preference of objects picked up should reflect the preference in the object fixated which is here modulated by cross-modal feature associations between colour and tactile features.

Here, the system behaviour is measured in terms of the objects picked up and relocated. We recorded only the first 25 reach and grasp actions (see Figure 3). Two runs are presented. In the first run, the base line, there is spatial modulation only. Consequently, object fixation and reaching actions are driven by the saliency map only which shows a slight bias towards blue objects.

In the second run cross-modal feature association learning takes place. Hence, the association between the neutral colour classes red (R) and blue (B) and the tactile classes hard (HA) and soft (SO) are learned. Therefore, a bias of fixating and picking up hard objects can only be caused by the cross-modal association between the tactile and the colour classes; not by the colour classes itself.

The results are shown in Figure 3. The left diagram shows the run of spatial modulation and the right shows the results for cross-modal learning. For spatial modulation only we see that more blue than red objects are picked up. This correlates with the bias towards blue objects generated by the saliency map. If cross-modal learning takes place then more red than blue objects are picked up. Hence, the system has learned to associate the red objects with the "desired hard tactile feedback" which leads to a highlighting of red objects due to cross-modal feature association.

IV. Discussion

The experiments shown have demonstrated how visual search is determined by object features and how visual and non-visual object features determine the system attention in terms of fixation patterns. Fixation patterns of human subjects are associated with visual attention indicating the computational processes in the brain that solve the given task (8). The experiments also validate our approach towards the integration of reaching, grasping and active vision. Furthermore, the solution we offer for the binding of "what" and "where" data is proven to be robust and works under real-time constraints. This is worth highlighting because comprehensive models solving this problem in robotics are not yet established. Also, nowadays many models in brain and vision research are still tested on static image data only, offering little help or convincing models when it comes to an application of these models for active vision and dynamic scenes.

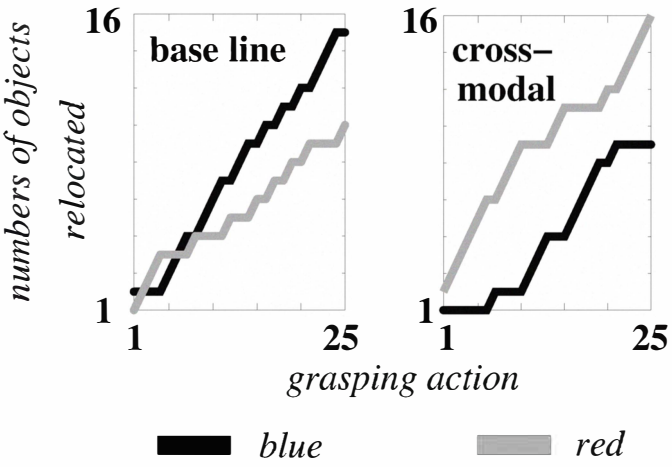

Fig. 3. Learning cross-modal feature association between colour and tactile feedback while picking up and relocation of objects.

However, this architecture provides the first milestone of our research towards a developmental learning framework for humanoid robots. In the following we will discuss our architecture from this perspective.

A. Computational domains and the sequence of developmental learning of hand-eye coordination in robots and humans

Considering the hand-eye coordination only, our architecture applies three computational domains: retinotopic, gaze and reach space. There is no external reference frame as it is often applied in robotics or assumed in brain research. However, we have an egocentric reference frame that provides the coordination of vision and reaching. Our egocentric reference frame is the gaze space. The usage of gaze space was not a superficial choice. It is based on our experiments on eye-saccade learning and was proposed and validated for robotics hand-eye coordination. There, we achieved hand-eye coordination without an external reference frame (7). Consequently, all the visual data are transformed from the retinotopic to the gaze space and from the gaze space into the reach space proving target coordinates for reach actions. Hence, gaze space is the central computational domain for vision and reaching. The transformations between the domains are done by two sensorimotor mappings: "eye-saccade mapping" and "gaze reach space mapping" (see Fig. 1). These mappings are learned before in separated experiments, but they should obviously subject of permanent learning or adaptation processes as it is the case for human beings. Interesting from a developmental perspective is the fact that the reach gaze space mapping can only be learned after the eye-saccade mapping is correctly learned. In other words, only if eye-saccades established the hand-eye coordination mapping can successfully learned.

This is in line with the developmental sequence of infants, who firstly establish eye-saccades and much later start to master hand-eye coordination (see (9) for a comprehensive discussion). Therefore we say, our architecture provides a framework for the study of developmental learning of hand-eye coordination without any external reference frame and

based only on the adaptation of two sensorimotor mappings. The organisation of the two sensorimotor mappings and the three computational domains is so general that it can easily be adapted to and be tested with other developmental learning strategies.

B. Hierarchies of cognitive competences for multimodal visual attention

Considering multimodal visual attention then our architecture establishes a hierarchy of competences. Without feature modulation, spatial memory and eye-saccade mapping then the active vision system can only generate uncoordinated camera movements. Providing the correct eye-saccade mapping, the system can fixate objects, but it is always attracted by the brightest stimuli. Adding the visual memory, we get an IOR mechanism and the system is able to fixate all the objects in its environment. With feature modulation the system can also highlight or inhibit specific features and therefor can be focused on objects useful for a specific task. And having finally task modulation as well then the system can change its focus depending on the currently given task.

Each competence is established on top of the other. This is also reflected in parts in Eq. 1, where starting with the original saliency value s the final activation value is modulated firstly by the spatial memory $(1-\frac{a}{M})$, while the spatial memory itself is modulated by the inhibition and excitation levels H and E.

This hierarchy of cognitive competencies is interesting from a robotic engineering point because it follows the principle of a subsumption architecture which is know to be robust and efficient. Furthermore, it is in line with the principle of "evolutionary refinement" outlined by M. Arbib as an principle that can be found biological systems and their "conceptual neural evolution" (10).

Finally, this hierarchy represents a sequence of developmental learning in human infants we try to model. Hence, once again the general organisation of the architecture provides a robotic reference for testing different approaches, strategies and models for acquiring the competence of multimodal visual attention in the characteristic stages of human development.

C. A cognitive robotics architecture

Many aspects of our architecture are not only inspired by the developmental sequence of human infants but also by brain models and behavioral studies. The problem of the egocentric reference frames and the transformation of visual and non-visual inputs into coordinated actions are ongoing research topics in brain research and neuroscience. This architecture can therefore be seen as an attempt to present a comprehensive model of active vision and object manipulation that integrates insights from brain research, vision research and developmental psychology. We call it "comprehensive" because the successful implementation on a robot system is a validation that also proves the solving of essential computational requirements with respect to feasibility, efficiency, scalability and real-time constrains.

V. Conclusion

We have presented an architecture that integrates active vision and simple object manipulation (reaching and grasping). As the experiments have shown the system is also able to integrate multimodal (visual and non-visual) features. We have further outlined how this architecture serves as a reference framework for the study of developmental learning; in particular, the staged behavioral competence learning of hand-eye coordination and reaching, as well as the hierarchical cognitive competence learning of multimodal visual attention. The architecture is therefore a promising robotic test bed for systematic investigations of mechanisms of developmental and open-ended learning.

Acknowledgment

This work was supported by the EC-FP7 projects IM-CLeVeR and ROSSI, and through UK EPSRC grant EP/C516303/1.

References

[1] M. Lungarella, G. Metta, R. Pfeifer, and G. Sandini, "Developmental robotics: a survey," *Connection Science*, vol. 15, no. 4, pp. 151–190, 2003.

[2] R. Pfeifer and J. Bongard, *How the Body Shapes the Way We Think: A New View of Intelligence*. MIT Press, 2006.

[3] M. Hülse and M. Lee, "Adaptation of coupled sensorimotor mappings: An investigation towards developmental learning of humanoids," in *11th International Conference on Simulation of Adaptive Behavior (SAB), LNAI 6226, Springer*, 2010, pp. 468–477.

[4] F. Chao, M. H. Lee, and J. J. Lee, "A developmental algorithm for ocular-motor coordination," *Robotics and Autonomous Systems*, vol. 58, pp. 239–248, 2010.

[5] M. Hülse, S. McBride, J. Law, and M. Lee, "Integration of active vision and reaching from a developmental robotics perspective," *IEEE Transactions on Autonomous Mental Development*, vol. 2, no. 4, pp. 355–367, 2010.

[6] L. Itti and C. Koch, "A saliency-based search mechanism for overt and covert shifts of visual attention," *Vision Research*, vol. 40, pp. 1489–1506, 2000.

[7] M. Hülse, S. McBride, and M. Lee, "Robotic hand-eye coordination without global reference: A biologically inspired learning scheme," in *In: Proc. Int. Conf. on Developmental Learing 2009, China, 2009, IEEE Catalog Number: CFP09294*, 2009.

[8] C. Rothkopf, D. Ballard, and M. M. Hayhoe, "Task and context determine where you look," *Journal of Vision*, vol. 7, pp. 1–20, 2007.

[9] J. Law, M. Lee, and M. Hülse, "Infant development sequences for shaping sensorimotor learning in humanoid robots," in *Proc. 10th International Conference on Epigenetic Robotics. Sweden*, 2010, pp. 65–72.

[10] M. A. Arbib, "Rana computatrix to human language: towards a computational neuroethology of language evolution," *Phil. Trans. R. Soc. Lond. A*, vol. 361, pp. 2345–2379, 2003.