# Integration of Active Vision and Reaching From a Developmental Robotics Perspective

Martin Hülse, *Member, IEEE*, Sebastian McBride, James Law, *Member, IEEE*, and Mark Lee

***Abstract*—Inspired by child development and brain research, we introduce a computational framework which integrates robotic active vision and reaching. Essential elements of this framework are sensorimotor mappings that link three different computational domains relating to visual data, gaze control, and reaching. The domain of gaze control is the central computational substrate that provides, first, a systematic visual search and, second, the transformation of visual data into coordinates for potential reach actions. In this respect, the representation of object locations emerges from the combination of sensorimotor mappings. The framework is tested in the form of two different architectures that perform visually guided reaching. Systematic experiments demonstrate how visual search influences reaching accuracy. The results of these experiments are discussed with respect to providing a reference architecture for developmental learning in humanoid robot systems.**

***Index Terms*—Biologically inspired robot architectures, developmental robotics, robotic active vision and reaching.**

## I. Introduction

ENGINEERING robot systems that interact autonomously, purposefully, and safely within our daily environment requires overcoming a number of challenges. One of the major challenges is the integration of different sensor and motor modalities, and the coordination of the many degrees of freedom (DOF) intrinsic to complex robotic systems. Historically, learning and adaptation processes have been extremely valuable in both mastering the aforementioned complexity, but also in deriving optimal parameter settings. Such processes of self-calibration and continuous adaptation also have the added advantage of allowing robotic systems to deal with changing and/or unconstrained environmental conditions.

Learning and adaptation, however, are not the final solution to the problem of efficient coordination of high dimensional systems. The reason being that any adaptation process will always perform poorly if applied to large parameter spaces. To overcome this issue, additional strategies of task decomposition, and the application of time regimes within the incremental learning process, must also be put in place. In other words, only by building up task complexity in incremental stages, such that all competences do not need to be learned at the same time, can modalities be accurately and reliably linked together.

There are many ways in which such a sequential paradigm of task composition and learning rate can be set up, but potentially the most productive strategy, especially with regard to humanoid robotics, is to adhere to the well defined and highly structured stages of child development. This approach also has the additional advantage of mimicking specific human characteristics, namely body shape, and actuator and sensor modalities, attributes consistently sought after within humanoid robots. Furthermore, based on the premise that incremental learning in infants occurs largely as a result of interaction with adults, it is possible to organise the learning process in such a way that it involves human-robot interaction, thereby facilitating the development of the humanoid robot through intuitively guided non-robotic-experts.

The particular way infants develop within the first months after birth might also indicate the way internal representations of the world are sequentially organized from a brain development perspective. For instance, if competence A is always developed before competence B, it may be that A is a prerequisite for B. It might also be that B can only be achieved if A is already present, or that B is in fact the result of the modulation of A (here, we understand behavior modulation to be equivalent to evolutionary refinement [1]). How we understand the order of occurrence of competencies in infants in the context of what is known about developing brain architectures might also give additional insights into how sensorimotor qualities can be represented. In this sense, combined findings in brain research and developmental psychology might provide a coherent guideline for the engineering of architectures that enable a humanoid, or similar, robot system to develop behavioral and cognitive capabilities analogous to humans.

This study, therefore, presents a conceptual framework for anthropomorphic robots that is inspired by findings in brain research, child development, and our own experiments on sensorimotor learning for autonomous robots. We present two computational architectures that implement this framework on a robot system solving a visually guided reaching task. The experiments with the robot system demonstrate the performance and requirements for visual search and robot reaching which, in turn, we use to evaluate our framework and the architectures with respect to our ultimate goal: an autonomously learning humanoid robot system that develops in a similar way to human infants during the first months of life.

## II. Overview and Motivation

The two computational architectures we present in this study allow a robotic active vision system, equipped with an arm, to perform visually guided reaching [Fig. 1 (Bottom)]. Thus,

Manuscript received March 01, 2010; revised August 23, 2010; accepted September 17, 2010. Date of publication September 30, 2010; date of current version December 10, 2010. This work was supported by the EC-FP7 Projects IM-CLeVeR and ROSSI, and through UK EPSRC Grant EP/C516303/1.

The authors are with the Department of Computer Science, Aberystwyth University, Aberystwyth, SY23 3DB, U.K. (e-mail: msh@aber.ac.uk; sdm@aber.ac.uk; jxl@aber.ac.uk; mhl@aber.ac.uk).

Digital Object Identifier 10.1109/TAMD.2010.2081667

1943-0604/$26.00 © 2010 IEEE

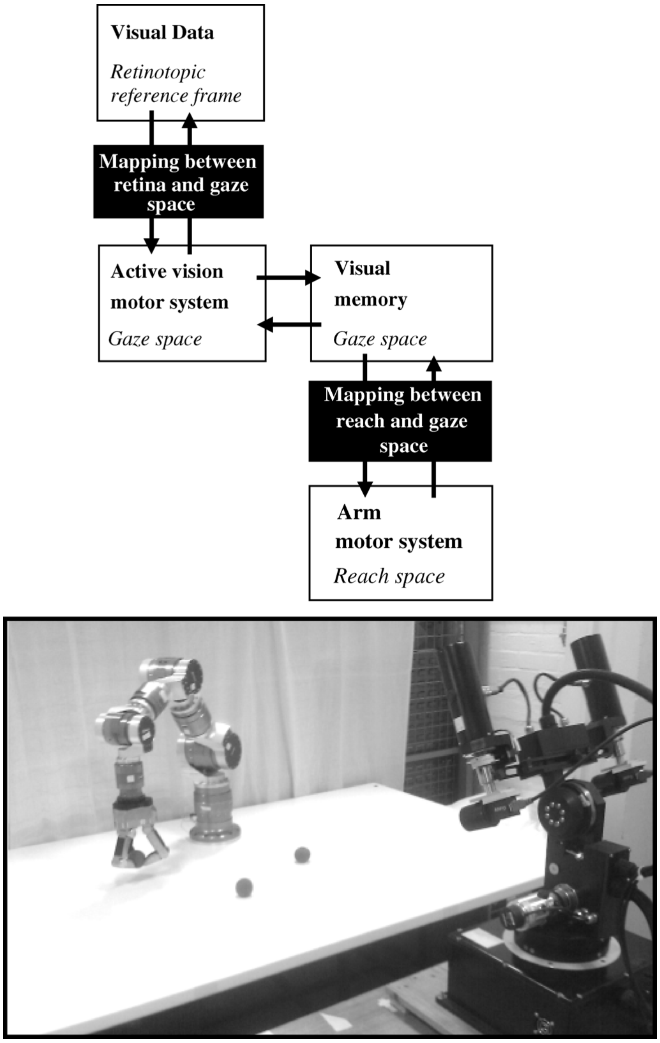
Visual Data
Retinotopic reference frame
Mapping between retina and gaze space
Active vision motor system
Gaze space
Visual memory
Gaze space
Mapping between reach and gaze space
Arm motor system
Reach space

Fig. 1. Top: Schema of the computational architecture integrating visual search and reaching competence. Bottom: Robotic system and scenario. An active vision system observes the objects on the table, which can be picked-up by a manipulator.

they are working examples for the integration of active vision and reaching competences. In fact, the two architectures are the result of a series of experiments on sensorimotor learning of eye-saccades [2], visual search [3], and hand–eye coordination [4], [5]. In Fig. 1 (Top), the general computational schema that both architectures instantiate is illustrated. Both operate in three different computational domains, the retinotopic reference frame, the gaze, and the reach space. All visual data are represented and processed in a retinotopic reference frame which is determined by the resolution and dimension of the image data that the camera delivers. The range of absolute motor positions of the active vision system defines the gaze space, where the latter represents the possible camera orientations. Finally, the reach space is defined by the robot arm coordinates that determine the points in space that the arm can reach to.

The three different domains are linked together by two mappings: the sensorimotor mapping for eye-saccades and the mapping between gaze and reach space. Within this experimental setup these mappings are already learned and fixed.

The last core element of our architectures is the visual memory, which stores motor configurations of the active vision system. Thus, it is represented in the gaze space. Two main functions are provided by the visual memory. First, it modulates visual search and therefore is essential to enable the system to perform a systematic visual scan of the environment, and second, its content represents potential reach targets (since the entries in the gaze space can be directly mapped into the reach space).

In summary, the two mappings provide the transformation of visual data into reach coordinates. This is only possible via the gaze space. In addition, visual search is also mediated by the visual memory and thus, the gaze space is the central element of this framework.

The motivation of this "gaze space centered approach" is threefold; first, it has a strong biological grounding based on the neurophysiological literature [6] (see Section II-A for further discussion). Second, the gaze space provides the global reference frame which is needed for a robust visual search by a robotic active vision system. Any retinotopic reference frame of an active vision system can only be local and cannot directly provide a global reference that allows robust visual search (see [7] for a detailed discussion). Third, the involvement of the gaze space instantiates a specific time line of staged competence learning that adheres closely to what is observed during infant development, thus providing additional biological validity (see Section II-B for further discussion). In particular, the mapping for eye saccades must be learned first before a robust and stable visual search competence can emerge, whilst only if visual search is performed successfully can a coherent hand–eye coordination be established to allow successful visually guided reaching. The sequence of competences can, therefore, only be achieved through the following stages:

1) eye-saccades;
2) visual search;
3) hand–eye coordination.

This leads to the last competence: visually guided reaching. Within the actual developing biological system, such discrete stages of competence do not exist, but rather the system progresses through overlapping parallel development. Such a synchronous learning process is currently being explored [8] and may be integrated into the overall architecture at a later date.

### A. Brain Areas Involved in Visual Search and Reaching

Intrinsic to any active vision system is the process of saccade, the voluntary or reactive movement of the vision system to bring selected parts of the visual scene into higher resolution. The biological structure responsible for this process is the superior colliculus; visual information passes from the superficial to the deep layer of this structure to elicit the correct sequence of oculomotor neurons and an accurate saccade [9]. This mapping process that links coordinates within the peripheral vision to the correct motor response establishes itself during the first seven months of infant development [10]. Infants begin by making highly variable primary saccades, followed by up to five secondary hypometric saccades, before bringing the image onto the fovea. By the seventh month, the mapping processes have

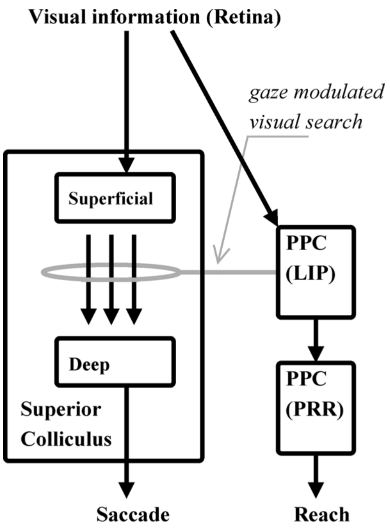

Fig. 2. Primary brain structures involved in visually guided reach: posterior parietal cortex (PPC); parietal reach region (PRR); lateral intraperietal area (LIP); medial intraperietal area (MIP).

produced a much more accurate primary saccade with only two or three secondary saccades required for accurate foveation.

Not saccading to an object previously saccaded to is also a critical attribute of the visual system, especially in the context of visual search. This is referred to as inhibition of return (IOR) and essentially refers to the suppression of stimuli (object and events) processing where those stimuli have previously (and recently) been the focus of spatial attention. In this sense, it forms the basis of attentional (and thus, visual) bias towards novel objects. Although the neural mechanism underpinning IOR is not completely understood, it is well established that the dorsal frontoparietal network, including frontal eye field (FEF) and superior parietal cortex, are the primary structures mediating its control. These are some of the many modulatory and affecting structures of the deep superior colliculus (optic tectum in nonmammals), the primary motor structure controlling saccade.

Although visual information from the retina starts at the superficial superior colliculus, and there are direct connections between the superior and deep layers [9], the former cannot elicit saccade directly [11]. This information has to be subsequently processed by a number of cortical and subcortical structures that place it: 1) in context of attentional bias within egocentric saliency maps (posterior parietal cortex) [12]; 2) the aforementioned IOR [13]; 3) overriding voluntary saccades (frontal eye fields) [9]; and 4) basal ganglia action selection [14]. Thus biologically, there exists a highly developed, context specific method for facilitating the most appropriate saccade as a form of attention selection. One of the main problems to overcome in constructing an IOR system is the accurate mapping of the retinotopic space to the global egocentric space, i.e., foveated objects within a retinotopic map must be logged within a global egocentric map to allow subsequent comparison with peripheral retinotopic information. The lateral intraparietal (LIP) area is the primary candidate brain region for this process given its position in modulating the transfer of visual information from superficial to deep superior colliculus.

As well as holding information about what objects have been saccaded to, the LIP region is also considered to contain information on task-specific object saliencies within the egocentric space [12]. The relay of this information to the medial intraparietal region (MIP) is the considered starting point for task-specific motor actions and, thus, LIP is central and critical to the saccade-reach process. It is these combined features of LIP that are the primary inspiration for the "gaze space centered approach" (Fig. 2).

In terms of eliciting a motor action, MIP transfers reach information to the premotor cortex in order to generate the movement vector that will take hand to target. Target location within MIP has been reported to be in eye-centered coordinates with gain modulation in relation to proprioceptive hand position [15]. The key to deriving a movement vector is to establish a common reference frame and this can either be done by: 1) coding hand position in eye-centered coordinates; or 2) coding the target in body-centered coordinates.

Using single neuron cell recordings within discrete regions of the posterior parietal cortex during different situations of reach performance, it appears that the parietal reach region codes proprioceptive information about hand position in eye-centered coordinates and that the movement vector is derived from simple subtraction of hand location from target position [16]. Developmentally, accurate reach does not take place until around four months [17] and this correlates with the known development of LIP and MIP structures within infants [18].

### *B. Infant Development*

Development in the human infant is restricted by a series of constraints, which restrain the infant's action repertoire and sensing capabilities. Initially, these constraints reduce the perceived complexity of the environment and limit interaction, providing a scaffold which helps the infant to make sense of the world [19], [20]. These constraints are then gradually eased, or lifted, allowing the infant to advance into a new stage of development [20]. In essence, such constraints prevent the infant from learning to run before it can walk.

In this section, we describe the process of learning to saccade, search, and reach, as it occurs in the human infant. Infants develop at different rates, and there is no set order or time line for the appropriation of skills, although the onset of some actions do commonly precede others. Developmental psychology supports the theory that early motor skills are related to perceptive and cognitive development [21], which varies from child to child.

*1) Saccades:* The eyes of the fetus can be seen to move in the womb from 18 weeks after conception, although the eyes stay closed until week 26 [22]. At birth, the eyes are the most controlled of the infant motor abilities, perhaps due to the lack of resistance in the socket. The neonate also has relatively poor vision, although saccades to what it can see are remarkably accurate. It is interesting to note that the newborn can saccade to stimuli at all, since in the womb there will be little, if any, visual stimuli.

Initially, the neonate can only focus at a distance of around 21 cm, which relates to the distance to the mother's face when held [22]. Color perception is coarse [23], but with similar categorization to adults by the second month [24], and the seen image is diffuse with a lack of clarity in the center of the visual field [25]. Perhaps due to these restrictions, the neonate is most attracted to diffuse lights and colors, and moving objects within its focal range [26].

Eye saccades in the neonate are relatively few in number. They are initiated in response to visual stimuli in the periphery of the visual field, but also in response to sounds [27]. Newborns tend to fixate on a single stimuli, but may be distracted by a sufficient peripheral stimuli. They are more likely to saccade to near objects than far ones, and to saccade horizontally, rather than vertically [28]. Although remarkably effective, these early saccades are not as accurate or rapid in onset as the adult variety, taking up to two seconds to trigger and often requiring several saccades to fixate the target [10], [28]. The mature form of saccade does not develop until after seven weeks of age, although this is sufficient for the young infant's requirements [29].

During the second month, the infant's ability to focus continues to improve, and acuity undergoes its greatest improvement [25]. The field of view also increases from around 20 degrees at six weeks to 40 degrees at 10 weeks [30]. The frequency of eye saccades increases with the majority of fixations now focused within objects [31]. The accuracy of saccades continues to improve and, correspondingly, the number of saccades required to fixate reduces, with near adult ability observed at seven months [10].

*2) Visual Search:* There is disagreement as to whether neonates are capable of visual search. Tronick and Clanton [32] interpreted infants' saccades in the absence of visual stimuli as visual search, although this was disputed by Salapatek [33] who suggested stimuli unobserved by the experimenters may have triggered the saccades. Further evidence provided by Haith [34] indicates newborns may be capable of visual search.

During the first month after birth, infants tend to fixate on object edges, rather than internal features [31]. Milewski [35] has shown that, as well as ignoring internal features at this age, young infants are incapable of remembering anything about them. By the second month searches are similar to adult levels, with more saccades and the majority of fixations now focused within objects [31]. The infant can follow moving stimuli to mid-line [36], although it shows very little head movement during gaze shifts of up to 30 degrees amplitude [37].

In the third month, the infant moves the head to assist visual search, making head movements about 25% of the time for 10 degree gaze shifts and all of the time for 30 degree gaze shifts [37]. Search continues to improve until the sixth month, by which time the infant is visually insatiable, moving head and eyes to search for, and fixate on, novel stimuli [26].

*3) Reaching:* In the womb, the limbs of the fetus demonstrate movement around eight weeks, with hand–face contact and flexing of fingers observed at week 10 [38]. However, at this age the brain is not yet developed enough for these actions to be intentional. General arm and finger movements appear to be refined over the following weeks. These movements are the result of lower brain functions and occur less frequently as neural myelination takes place between 18 and 24 weeks.

Hand-to-mouth movements continue after birth, and were initially also considered to be purely reflexive [39], [40]. However, more recent studies by Butterworth and Hopkins [41] and Rochat [42] have shown that these actions may actually be intentional. Trevarthen [43] and Von Hofsten [44] have shown that, provided they have adequate support, newborns can make general reaching movements toward visual stimuli. The neonate has a reflex grasp, and so may even close its fingers around objects if it makes contact. Nevertheless, this type of reaching is ballistic, and uncoordinated; newborns do not attempt to adjust their reaching pattern during the reaching movement, and simply withdraw if they are unsuccessful. Although the reaching is goal directed, and initiated by a visual stimulus, it is not guided by visual feedback. For this reason, this early reaching is called *visually elicited reaching*. It disappears during the first seven weeks after birth to be replaced by the onset of more deliberative reaching [45].

The infant continues to improve the control of its limbs with the first successful reaching, where the hand regularly makes contact with the target, occurring at around three months. This coincides with a shift from one- to two-handed reaching [36], [46]. Reaches consist of a series of jerky movements, as the infant attempts to control the motion by altering the limb stiffness [47]. By this stage the infant can grasp and reach at will, but actions are still visually and tactually elicited [22].

Around four to five months after birth, the infant begins to bring reaching and grasping together under the control of visual feedback [48]. Although currently requiring multiple movements to reach to an object, the infant will learn to make smoother, more continuous, movements over the coming months. Hand–eye coordination is well developed by nine months, with the infant being able to reliably manipulate objects and pass them from hand to hand [26].

Roughly summarizing this description of the first months of infant development, it seems that competence learning is achieved in part successively, but also in parallel. The foundations for saccade, visual search, and reaching are all in place at birth, to some extent, however, they are not sufficient to perform the task of visually guided reaching, and are lacking additional competencies required to do so (such as intent). All three abilities improve over the first months of life, although they may also deteriorate before doing so (as is the case for reaching).

Of all three abilities, the saccade is most developed at birth, and the fastest to improve, followed by visual search, and finally reaching. The stages of development coincide to ensure that the ability to saccade is reasonably developed by the time visual search commences, and that both are well established before there is sufficient arm control for purposeful reaching. All three abilities must be in place before visually guided reaching can be performed, and so the onset of that final stage is constrained by the level of refinement of the constituent parts.

Although all three abilities develop consecutively, there is an underlying series of constraints that require saccade function to precede visual search, to precede reach. We use this sequence to inform the development of our robotic architectures and our gaze space centered framework towards visually guided reaching: eye-saccades first, followed by visual search, finally followed by hand–eye coordination, which leads to visually guided reaching.

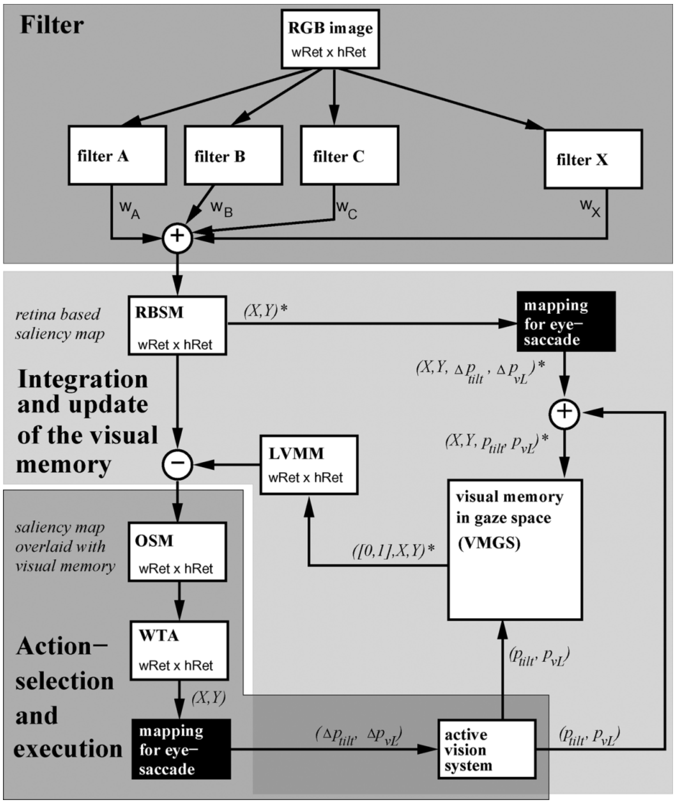
Filter
RGB image
wRet x hRet
filter A
filter B
filter C
filter X
$w_A$
$w_B$
$w_C$
$w_X$
retina based saliency map
RBSM
wRet x hRet
(X,Y)*
mapping for eye−saccade
$(X,Y, \Delta p_{tilt}, \Delta p_{vL})^*$
Integration and update of the visual memory
LVMM
wRet x hRet
$(X,Y, p_{tilt}, p_{vL})^*$
visual memory in gaze space (VMGS)
([0,1],X,Y)*
saliency map overlaid with visual memory
OSM
wRet x hRet
WTA
wRet x hRet
$(p_{tilt}, p_{vL})$
Action−selection and execution
(X,Y)
mapping for eye−saccade
$(\Delta p_{tilt}, \Delta p_{vL})$
active vision system
$(p_{tilt}, p_{vL})$

Fig. 3. Computational architecture $A_{\mathcal{R}}$.

### C. Specific Aims of This Study

In this work, we introduce two architectures that instantiate the above described gaze space centered approach towards visually guided reaching. Both architectures differ in the way visual search is modulated. We will demonstrate that this difference has an impact on the achieved accuracy of reaching. The experimental results we provide will allow an evaluation of these architectures with respect to their future use as reference architectures for developmental learning in humanoid, and similar, robot systems.

## III. Methods

The active vision system consists of two cameras (both provide RGB $1032 \times 778$ image data) mounted on a motorized pan-tilt-verge unit (Fig. 1). Here, only one camera and two degrees of freedom (DOF) are used: the left camera verge movement, and tilt. Each motor is controlled by determining its absolute target position, or the change of the current position, given in radians (rad). The use of only one camera and two degrees of freedom was sufficient in the context of the current architectures in order to get the same reaching precision when the second camera was involved [5]. It was not necessary therefore to use the second camera system. Thus, the active vision system configuration is fully determined by the absolute motor positions of the tilt and left verge axis, $(p_{\text{tilt}}, p_{vL})$. The absolute positions of these two parameters define the *gaze space*.

The active vision system is oriented towards a table where colored objects (balls) are placed. A robotic manipulator is mounted on the same table in order to pick-up and move these objects. The objects do not change their location. Hence, the objects on the table provide a static scenario for the active vision system.

The robot arm and hand systems (SCHUNK GmbH & Co. KG) have seven DOF each, but only five DOF of the arm are used in order to place the robot hand at certain positions on the table. The hand system has three fingers, each with two segments equipped with a pressure sensitive sensor pad. Since the objects used in this scenario have the same shape and consistency, only one grasping procedure was applied in this experiment. The control of the grasping is out of the scope of this paper.

Since the objects are only located on a table, the domain of the reach movement, referred to here as *reach space*, is represented as a 2-dimensional polar coordinate system. Taking the base of the arm as a reference, a table location is fully determined by the distance $d$ (cm) and the planar angle of the arm $\alpha$ (rad). The inverse kinematic mapping between the 2-dimensional reach space and the 5-dimensional joint space of the arm is solved analytically, and is not described here. It is important to note that arm control only places the hand on the table with respect to a given distance and relative angle $(d, \alpha)$. The actual table space the system is operating in is defined by the range of distances $d$ and angle $\alpha$, here: $-1.4 \leq \alpha \leq 1.4$(rad) and $33 \leq d \leq 59$(cm).

## IV. Two Computational Architectures for Gaze Modulation

In the following, we introduce two computational architectures for gaze-modulated visual search. Both architectures use mapping processes to facilitate saccade action, where $(X, Y)$ coordinates of the local retinotopic image data are transformed into motor position changes $(\Delta p_{\text{tilt}}, \Delta p_{vL})$ given in *rad*. The execution of these motor position changes drives the camera in such a way that the corresponding stimulus at the $(X, Y)$ coordinates end up in the fovea, i.e., the image center. The actual saccade mappings can either be learned [49], [50] or manually designed. The latter was employed for this study.

The central element of both architectures is the visual memory, which stores the absolute motor configuration of the active vision system $(p_{\text{tilt}}, p_{vL})$ after the execution of a successful saccade. A saccade is successful if the object is driven into the central region of the image; the assumed location of the fovea. The domain of the visual memory is the gaze space and is referred to as VMGS (visual memory in gaze space).

We now introduce two very distinct strategies for how visual search can be modulated by the content of the visual memory (VMGS) to generate an IOR mechanism. In the first architecture $A_{\mathcal{R}}$, Figs. 3 and 4, the suppression of stimuli which the system has already saccaded to is performed in the domain of the local retinotopic reference frame. Thus, the inhibition of return is operating in the local retina space. In the second architecture $A_{\mathcal{G}}$, Fig. 5, stimuli that the system has already saccaded to are suppressed in the gaze space. As a consequence, the final action selection process for eye-saccades is performed in the

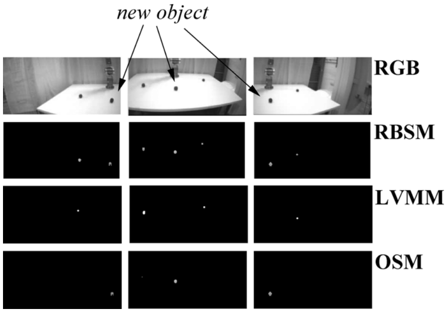

Fig. 4. Particular system states of architecture $A_{\mathcal{R}}$ for different camera positions after a new object is placed on the table. The new object is not yet present in the visual memory, therefore it is the only stimuli in the OSM. The stimuli representing the old objects are present in LVMM, which inhibits their emergence in the OSM. See text for details.

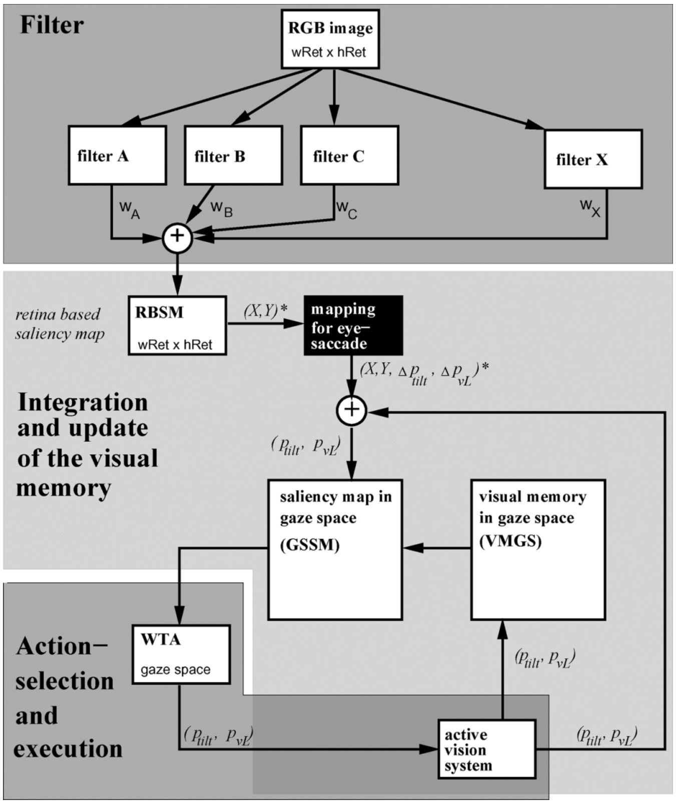

Fig. 5. Computational architecture $A_{\mathcal{G}}$.

global gaze-space for $A_{\mathcal{G}}$. The following sections provide more detailed descriptions of the two architectures.

### A. Eye-Saccade Action Selection in Retina Space

The overall computational architecture of $A_{\mathcal{R}}$ consists of three main functional stages that implement: 1) filtering of image data; 2) action selection and execution; and 3) the processing of the *visual memory* VMGS.

A color filtering process on the current camera image data generates a saliency map referred to as the retina-based saliency map (RBSM). The dimensions of the RBSM are determined by the width and height of the camera images ($\mathrm{wRet} \times \mathrm{hRet}$). Each stimuli in RBSM is represented by a nonzero entry of the corresponding $(X, Y)$ coordinates. Due to the previously learned eye-saccade mapping, each $(X, Y)$ coordinate of a nonzero entry in the RBSM elicits a corresponding motor change $(\Delta p_{\mathrm{tilt}}, \Delta p_{vL})$ for a successful saccade. Together with the current absolute motor positions (delivered by the active vision system), this produces, for each nonzero $(X, Y)$ coordinate, the expected absolute motor positions of the vision system if a saccade towards this stimulus was executed. This is expressed in Fig. 3 in the form of $(X, Y, p_{\mathrm{tilt}}, p_{vL})^*$, which refers to a list of all nonzero $(X, Y)$ coordinates and their expected final absolute motor configurations after the corresponding saccade. These potential motor configurations are tested against the current entries of the visual memory VMGS. If the potential absolute motor configuration is present in VMGS then the corresponding $(X, Y)$ coordinate is labeled 1, otherwise 0. In this way, we get a new list $([0/1], X, Y)^*$ of all nonzero $(X, Y)$ coordinates in the RBSM which are labeled according to their presence in the VMGS. This list can be transformed into a local visual memory map (LVMM), having the same dimensions as RBSM. In contrast to RBSM, the nonzero entries in LVMM represent the stimuli the system has already saccaded to. Hence, the subtraction of RBSM from LVMM will generate a new map, overlaid saliency map (OSM), which contains only the stimuli which have not yet been saccaded to by the active vision system.

The OSM is fed into the action selection process, which is implemented as a winner-take-all (WTA) process. If the subsequent saccade execution is successful, the final configuration $(p_{\mathrm{tilt}}, p_{vL})$ is stored as a new entry in the visual memory (VMGS).

The usage of the gaze space as a domain for representing the visual memory provides the globally acting IOR. Fig. 4 represents the image data (RGB) as well as the RBSM, LVMM, and OSM data for different camera positions. The nonblack entries represent stimuli, or nonzero activations, and black pixel values indicate zero activation values. In this particular scenario we started with two objects on the table. After the active vision system stored them in its visual memory (by executing saccades towards them) the saccade process was turned off and a new object was placed on the table. One can clearly see, for arbitrary camera positions, that there is only one stimuli present in the OSM. The LVMM, however, contains stimuli (one or two) which correspond to the objects the system has already stored in the visual memory. Thus, in any camera position, only the new object is fed into the action selection process for the saccadic eye-movement. Notice that even if the old objects fall out of the visual field (left and right image in Fig. 4), as soon as they are back the system will inhibit them again. The IOR thus acts locally on the current image input but is stored globally in the gaze space.

### B. Eye-Saccade Action Selection in Gaze Space

The second architecture $A_{\mathcal{G}}$ (Fig. 5) has a structure similar to $A_{\mathcal{R}}$ in that there are three main functional parts: image data filtering, action selection, and visual memory. The processing between the two architectures differs after the generation of the RBSM in that now all stimuli in the RBSM are mapped into the gaze space. Hence, instead of a retina-based saliency map, we have now a gaze-based saliency map (GSSM). The

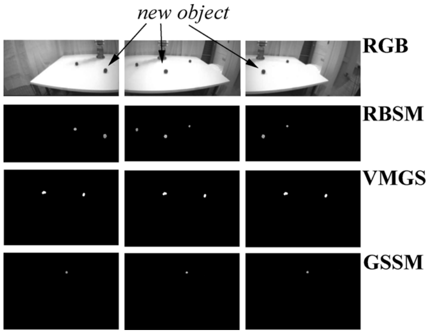

Fig. 6. Particular system states of architecture $A_{\mathcal{G}}$ for different camera positions after a new object is placed on the table. The new object is not yet present in the visual memory VMGS. Since VMGS directly inhibits the gaze space based saliency map, only one stimulus is present in GSSM. See text for details.

process of transformation is the same as for architecture $A_{\mathcal{R}}$. For each stimuli in RBSM the expected final absolute motor configuration of the potential saccade is derived but, instead of testing each $(p_{\text{tilt}}, p_{vL})$-configuration for each potential saccade against the visual memory (VMGS), they all are stored in the GSSM.

The current GSSM is fed into the same action selection process (WTA) with the outputs as absolute target positions $(p_{\text{tilt}}, p_{vL})$. If the movement of the camera to this target position represents a successful saccade, it will again be stored in the visual memory VMGS.

With respect to the IOR mechanisms, the VMGS can directly inhibit the GSSM because both are represented in the same domain. Thus, the IOR mechanism operates exclusively in the gaze space.

Here again we have plotted the image data and resulting VMGS and GSSM configurations for different camera positions (Fig. 6). As in the previous scenario, we started with two objects on the table. After the two objects were stored by the system in the VMGS, the saccade execution was deactivated and a new object was placed on the table. Inhibition by the visual memory, VMGS, means that only the stimulus of the new object emerges in the GSSM. Since the GSSM represents the data fed into the action selection process, the next saccade would lead to the fixation of the new object.

## V. Integration of Reach Competence

As visual memory, VMGS, is the central element for the visual search, so it is for reaching. For both architectures it is the VMGS which links reaching and visual search (Fig. 7). The current content of the VMGS is fed into an action selection process (again a winner-takes-all strategy). The selected target position in gaze space $(p_{\text{tilt}}, p_{vL})$ is translated into the corresponding coordinates in the reach space $(d, \alpha)$. The learning of such a mapping between gaze and reach space is demonstrated in [4], [5]. It is a model-free case-based learning method which allows faster learning when compared with artificial neural networks [49]. However, the learning experiments did not only generated the mapping applied to these architectures here. It also provided a base-line for the expected average reaching accuracy of this robotic system, which is $(2.00 \pm 1.0 \text{ cm})$ [5].

Furthermore, it is important to notice that the applied mapping is bidirectional. Thus, after successfully grasping an object at the given target position, this position in reach space can be mapped back into the gaze space. These data are used to remove the corresponding entries in the visual memory VMGS. This has two important consequences. F irst, the corresponding object location is not inhibited anymore by the VMGS and therefore new objects which might be placed on this position can immediately be detected by the robot system. Second, the next action selection process initiating reaching does not involve stimuli representing the object which has been picked-up already. Hence, it guarantees that the robot reaches only to objects the system has not yet reached to.

In this setup, the grasped objects were not put back into the work area. In other words, one object after the other is removed by the robot. The success of a grasp is indicated by the pressure-sensitive finger pads of the robot hand.

## VI. Experiments

In implementing the two architectures $A_{\mathcal{R}}$ and $A_{\mathcal{G}}$ for our robotic system, we have to consider one essential parameter, called $\epsilon$. The value of $\epsilon$ defines the maximal distance between two points in the gaze space such that they can be considered to be the same object. This is necessary due to the noise associated with the variation in active vision configurations for saccades towards the same object. $\epsilon$, therefore, defines a neighborhood for a specific gaze space configuration to compensate for this noise effect.

Obviously, $\epsilon$ will highly influence the number of elements stored in the visual memory, VMGS, and therefore the behavioral dynamics of the active vision system, which in turn determines the reaching performance.

In the following, we will present two sets of experiments. The first is focused on the visual search only where it is demonstrated how $\epsilon$ determines the number of saccades needed to scan a visual scene. In the second set of experiments, the performance of the visually driven reaching is tested for different object positions and $\epsilon$-values.

### A. Visual Search Only

Here we present experiments conducted to test the impact of $\epsilon$-values on the visual search. The reaching component is not involved. In each experiment three objects were placed on the table, similar to the scenario shown in Figs. 4 and 6. For each parameter setting, the visual memory VMGS was empty and the starting orientation of the camera was kept constant.

Saccades were measured via the output recording of the absolute positions of the verge motor. Fixating the objects on the table, the corresponding verge motor position values are very distinct ($\approx -0.5$, $\approx 0.2$, and $\approx 0.5$ rad) and allow us to derive the correlation between the state of the active vision system and the fixated object or eye-saccade execution.

Once a $(p_{\text{tilt}}, p_{vL})$ configuration of a successful saccade is stored in the VMGS, it remains there. Thus, the visual search will generate saccades to all the stimuli until the stored configurations in VMGS cover the whole scene, at which point the system will stop saccading. The metric of the $\epsilon$-parameter was

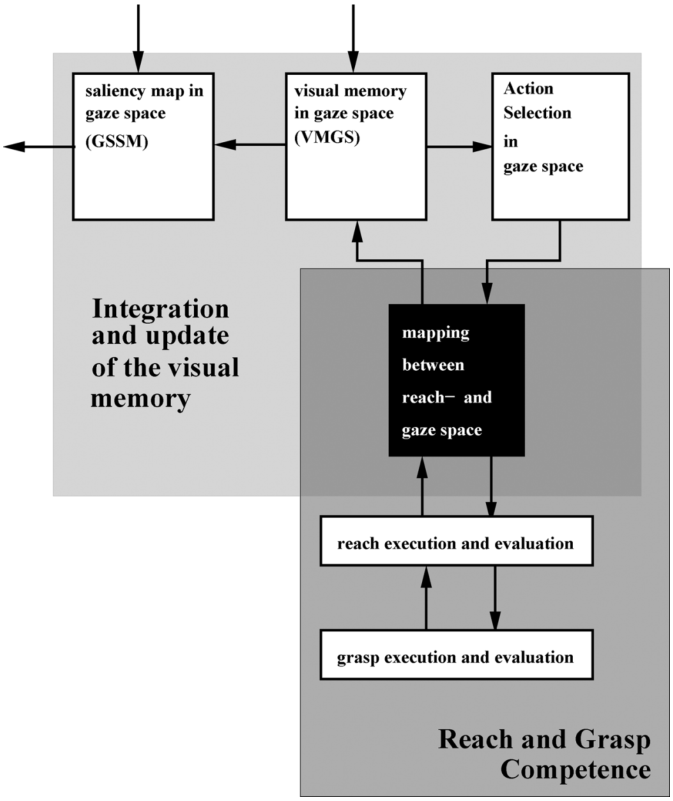
saliency map in gaze space (GSSM)
visual memory in gaze space (VMGS)
Action Selection in gaze space
Integration and update of the visual memory
mapping between reach− and gaze space
reach execution and evaluation
grasp execution and evaluation
Reach and Grasp Competence

Fig. 7. Integration of reach control for both architectures.

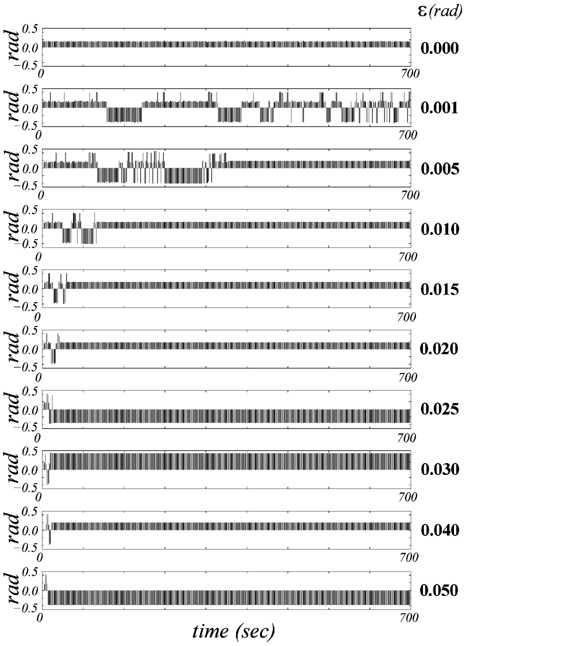

Fig. 8. Resulting verge positions for different $\epsilon$-values for architecture $A_{\mathcal{R}}$. (Data shown for the first 700 s.)

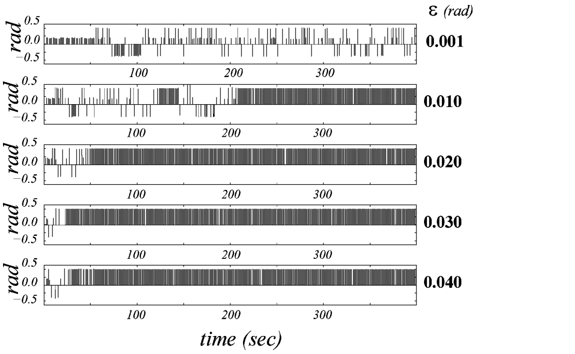

Fig. 9. Resulting verge positions for different nonzero $\epsilon$-values of architecture $A_{\mathcal{G}}$. (Data shown for the first 400 s.)

TABLE I
DURATION OF SACCADING PROCESS

| $\varepsilon$ (*rad*) | Time (sec.) until final saccade | |
|---|---|---|
| | $A_{\mathscr{R}}$ | $A_{\mathscr{G}}$ |
| 0.010 | 103 | 213 |
| 0.020 | 33 | 50 |
| 0.030 | 26 | 23 |
| 0.040 | 10 | 54 |

the Euclidean distance and each test run was conducted over 800 seconds.

A selection of runs are presented in Fig. 8 (architecture $A_{\mathcal{R}}$) and Fig. 9 ($A_{\mathcal{G}}$). The experiment with $\epsilon = 0.0$ (shown in Fig. 8) essentially illustrates system behavior without any inhibition mechanism. Here, the system remains in the same configuration apart from small fluctuations. After the camera has saccaded to the most salient stimulus, it remains in the same position since a neighborhood of zero results in no inhibition of nearby pixels generated by the same object. Theoretically, a large number of different saccades should finally lead to a coverage of the whole object stimulus. However, due to the limited precision of the active vision system, a target position might not be achieved by the system's actuators. In such a situation a close-to-zero-neighborhood will never lead to total inhibition of all the stimuli generated by one object of a reasonable size.

In general, however, the plots show that the larger the $\epsilon$-value the fewer saccades or $(p_{\text{tilt}}, p_{vL})$ configurations in VMGS were necessary to inhibit the stimuli generated by the objects. This was indicated by the time and number of saccades required until the active vision system stopped. Although a qualitatively similar trend of behavior was generated by architecture $A_{\mathcal{G}}$ (Fig. 9), the time taken to reach execution of the final saccade was generally longer compared to architecture $A_{\mathcal{R}}$ (see Table I for numerical values).

On the other hand, if $\epsilon$ is too small (here $0 < \epsilon \leq 0.001$), the saccade process does not stop because the actuators are not able to precisely drive into the gaze space configuration needed to cover all stimuli in GSSM ($A_{\mathcal{G}}$) and RBSM ($A_{\mathcal{G}}$), respectively.

## B. Evaluating the Visually Guided Reach

In this series of experiments, we have evaluated the accuracy of reaching after the complete scanning of a scene by the active vision system. This was done for both architectures, $A_{\mathcal{R}}$ and $A_{\mathcal{G}}$. Starting with an empty visual memory, VMGS, and $\epsilon \geq 0.005$ the system saccades towards three objects on the table. After the active vision system had stopped reaching was triggered, and the resulting target positions in reach space were compared with the actual object positions. The distance on the

TABLE II
AVERAGE REACH ERROR VALUES FOR ARCHITECTURES $A_{\mathcal{R}}$ AND $A_{\mathcal{G}}$

| | Object position in reach space | | $\varepsilon$ $(rad)$ | | | | | | Total average over object positions |
|---|---|---|---|---|---|---|---|---|---|
| | $d(cm)$ | $\alpha(rad)$ | 0.005 | 0.010 | 0.015 | 0.020 | 0.025 | 0.030 | |
| $A_{\mathscr{R}}$ | | | | | | | | | |
| | 33 | -1.4 | 2.97 ± 0.00 | 2.37 ± 0.70 | 2.26 ± 0.49 | 2.87 ± 0.32 | 1.87 ± 0.72 | 1.96 ± 0.46 | 2.38 ± 0.65 |
| | | 0.0 | 3.02 ± 0.00 | 3.08 ± 0.09 | 2.81 ± 0.88 | 3.01 ± 0.87 | 1.56 ± 0.55 | 2.89 ± 0.55 | 2.70 ± 0.78 |
| | | 1.4 | 4.16 ± 0.09 | 3.07 ± 0.98 | 2.56 ± 0.84 | 1.51 ± 0.68 | 2.12 ± 0.85 | 2.00 ± 0.00 | 2.61 ± 1.10 |
| | 59 | -1.4 | 2.46 ± 2.00 | 2.88 ± 2.14 | 3.76 ± 2.07 | 2.54 ± 2.21 | 1.16 ± 0.00 | 2.10 ± 1.27 | 2.48 ± 1.89 |
| | | 0.0 | 2.31 ± 0.00 | 2.31 ± 0.00 | 2.31 ± 0.00 | 2.29 ± 0.07 | 2.29 ± 0.07 | 2.24 ± 0.11 | 2.29 ± 0.06 |
| | | 1.4 | 7.05 ± 1.94 | 7.66 ± 0.00 | 6.44 ± 2.58 | 7.05 ± 1.94 | 6.44 ± 2.58 | 2.15 ± 1.94 | 6.13 ± 2.67 |
| | Total average over $\varepsilon$-values | | 3.94 ± 2.72 | 4.28 ± 2.72 | 4.17 ± 2.53 | 3.96 ± 2.76 | 3.29 ± 2.72 | 2.17 ± 1.29 | $\sum$3.10 ± 2.00 |
| $A_{\mathscr{G}}$ | | | | | | | | | |
| | 33 | -1.4 | 2.26 ± 0.49 | 2.64 ± 0.32 | 2.05 ± 0.32 | 2.15 ± 0.43 | 2.78 ± 0.61 | 2.46 ± 0.53 | 2.39 ± 0.53 |
| | | 0.0 | 2.25 ± 0.78 | 2.55 ± 0.65 | 1.97 ± 0.77 | 2.58 ± 0.64 | 2.69 ± 0.82 | 2.48 ± 1.22 | 2.42 ± 0.84 |
| | | 1.4 | 2.00 ± 0.00 | 2.83 ± 1.07 | 2.21 ± 0.65 | 2.00 ± 0.00 | 2.21 ± 0.65 | 3.57 ± 1.52 | 2.47 ± 0.99 |
| | 59 | -1.4 | 2.36 ± 0.87 | 2.50 ± 0.96 | 2.71 ± 1.44 | 2.48 ± 1.57 | 3.54 ± 1.59 | 3.90 ± 2.32 | 2.92 ± 1.59 |
| | | 0.0 | 2.25 ± 0.13 | 2.35 ± 0.16 | 2.29 ± 0.07 | 2.29 ± 0.07 | 2.31 ± 0.00 | 2.31 ± 0.00 | 2.30 ± 0.10 |
| | | 1.4 | 1.54 ± 0.00 | 2.15 ± 1.94 | 4.60 ± 3.23 | 7.66 ± 0.00 | 3.37 ± 2.96 | 1.54 ± 0.00 | 3.48 ± 2.87 |
| | Total average over $\varepsilon$-values | | 2.11 ± 0.57 | 2.50 ± 1.02 | 2.64 ± 1.71 | 3.19 ± 2.14 | 2.82 ± 1.48 | 2.71 ± 1.45 | $\sum$2.66 ± 1.51 |

All error values given in *cm*.

table between estimated target position and actual position are called *reaching error values*.

Reaching error values were collected in a systematic way by selecting specific object positions and $\epsilon$-values ($0.005 \leq \epsilon \leq 0.030$). For each configuration, ten visual search processes were run leading to ten error values. In such a way, a total of 360 reaching error values were collected for each architecture.

The average errors with respect to the $\epsilon$-values and the object positions are summarized in Table II (all values given in *cm*).

Comparing only the two architectures it turns out that $A_{\mathcal{G}}$ produces, on average, better estimations than $A_{\mathcal{R}}$ ( $2.66 \pm 1.51$ cm versus $3.10 \pm 2.00$ cm). However, both average errors are much higher compared to the average errors achieved while learning the applied mapping between gaze and reach space ($2.00 \pm 1.0$ cm), which indicates our base-line [5]. This base-line is determined by the whole robotic setup including specifics of the robotic hardware, the spatial configuration of the robot arm and the vision system, as well as the learning method. Changes in this setup might result in a different overall average reaching error indicating the lower bound of the reaching accuracy that can be expected for the visual search task.

Comparing the average of the reaching errors values with respect to the $\epsilon$-values one can see that the best estimations for architecture $A_{\mathcal{G}}$ are achieved for $\epsilon = 0.005$, the smallest value. In contrast to $A_{\mathcal{R}}$, where $\epsilon = 0.030$ (the largest value) led to the best estimation.

The data also indicate differences with respect to the object positions, but these differences are qualitatively very similar between the two architectures. Both architectures have their worst estimation results in the (59, 1.4) position. Very similar estimations results can also be found in the (59, 0.0) position. For all the positions of distance 33 cm, we can also see similar average errors. In summary, object position related error values were the same for both architectures, suggesting that the error was caused by other factors. For example, they may have been caused by the mapping applied or by the given light conditions.

TABLE III
OVERVIEW OF THE DISCUSSED APPROACHES TO VISUAL SEARCH

| Approach to visual search | Domain | | |
|---|---|---|---|
| | Action selection | Visual memory | IOR |
| $A_{\mathscr{R}}$ | RT | global GS | RT |
| $A_{\mathscr{G}}$ | local GS | global GS | global GS |

(RT, retinotopic reference frame; GS, gaze space reference frame)

## VII. DISCUSSION

### A. Interplay of Local and Global Domains for Visual Search

It is interesting to compare our two architectures, $A_{\mathcal{R}}$ and $A_{\mathcal{G}}$, with respect to the domains of the different processing tasks involved in the visual search, namely: 1) eye-saccade action selection; 2) visual memory; and 3) the domain the IOR is operating on, i.e., where the suppression of visual stimuli takes place.

The two architectures presented within this study make use of a global reference frame, the gaze space, in order to represent the visual memory. However, for architecture $A_{\mathcal{R}}$, eye-saccade action selection is done in the retina space (OSM in Fig. 3) which is also the domain for the suppression of stimuli (subtraction of LVMM from RBSM). Whereas in architecture $A_{\mathcal{G}}$, action selection is done in the local gaze space. It is referred to as local because the stimuli in GSSM are determined by the stimuli in RBSM, which represents the current visual input (retina space) only. The suppression of GSSM by VMGS, however, acts in the global domain of the gaze space.

Table III provides an overview of the domains involved. We see both architectures facilitate an interplay between local and global reference frames. This is possible because of the mapping between the local retina space and the global gaze space. As we have seen in architecture $A_{\mathcal{R}}$, this mapping can even be bidirectional, from RBSM (retina space) to VMGS (gaze space) to LVMM and OSM (both retina space). Although the mapping itself (mapping for eye saccades) is not bidirectional, it maps retina coordinates $(X, Y)$ to relative motor positions $\Delta p$. A mapping that transforms the absolute motor position of the

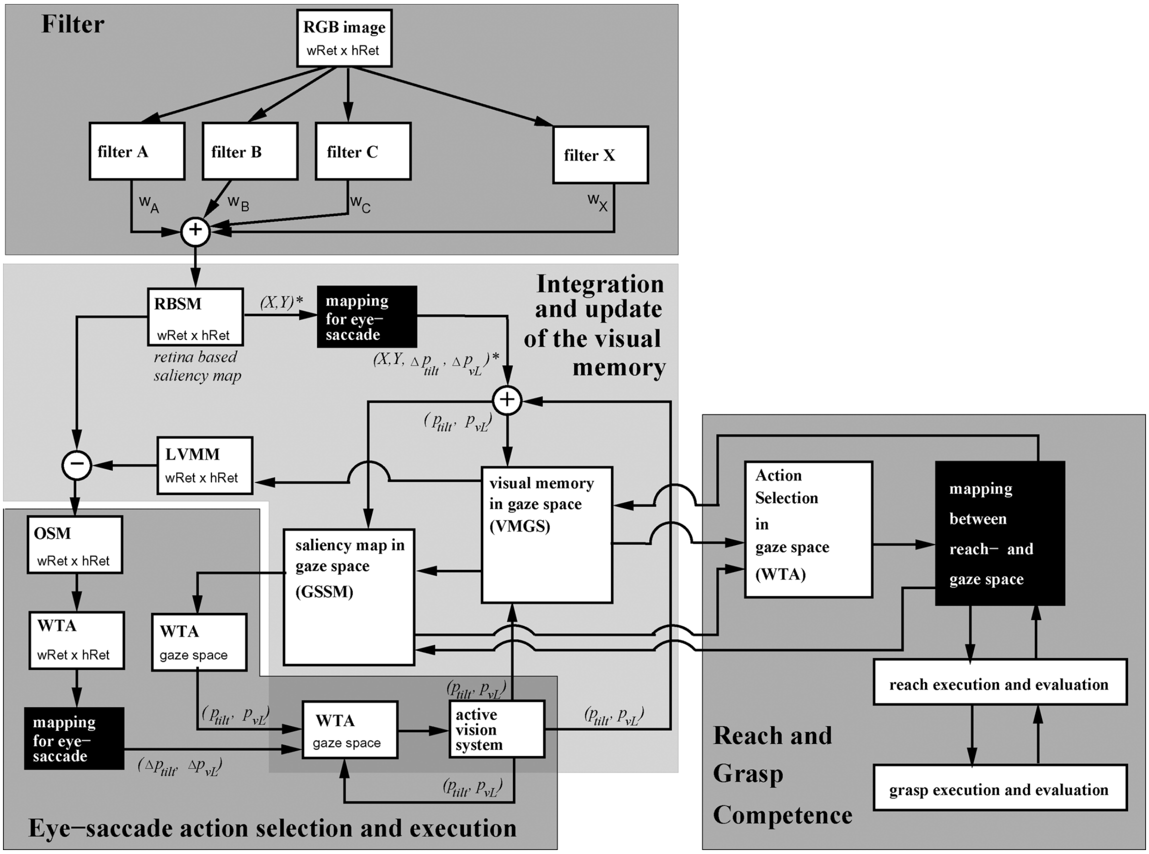
Filter
RGB image
wRet x hRet
filter A
filter B
filter C
filter X
$w_A$
$w_B$
$w_C$
$w_X$
+
Integration and update of the visual memory
RBSM
wRet x hRet
retina based saliency map
(X,Y)*
mapping for eye−saccade
$(X, Y, \Delta p_{tilt}, \Delta p_{vL})^*$
+
$(p_{tilt}, p_{vL})$
LVMM
wRet x hRet
−
visual memory in gaze space (VMGS)
OSM
wRet x hRet
saliency map in gaze space (GSSM)
WTA
wRet x hRet
WTA
gaze space
mapping for eye−saccade
$(p_{tilt}, p_{vL})$
$(\Delta p_{tilt}, \Delta p_{vL})$
WTA
gaze space
$(p_{tilt}, p_{vL})$
active vision system
$(p_{tilt}, p_{vL})$
$(p_{tilt}, p_{vL})$
Eye−saccade action selection and execution
Action Selection in gaze space (WTA)
mapping between reach− and gaze space
reach execution and evaluation
grasp execution and evaluation
Reach and Grasp Competence

Fig. 10. Combined architecture. See text for explanation.

active vision system $p$ into retina coordinates $(X, Y)$ does not exist. This is only achieved indirectly and through the actions performed by the active vision system.

### B. Visual Search Determines the Reaching Precision

For both architectures, the achieved accuracy of reaching is determined by the $\epsilon$-values, thus, by the modulation of the visual search by the visual memory. The results show that for $A_{\mathcal{R}}$ the best estimations are achieved for large $\epsilon$ values, while for $A_{\mathcal{G}}$ small $\epsilon$ values provided the best accuracy.

In general, the $\epsilon$ value relates to the number of saccades needed to cover all the stimuli caused by an object. Consequently, the larger the $\epsilon$-values the less the number of entries in the VMGS that represent a certain object. These data would suggest that $A_{\mathcal{R}}$ performs better if only a few saccades are made towards the objects, while vice versa for $A_{\mathcal{G}}$. However, such a conclusion might be misleading. In fact, it is not the number of saccades that determine the achieved accuracy, it is the "quality" of stored configurations in VMGS representing the saccades made towards the object. This "quality" results from the whole process of visual search involving the mapping between retina and gaze space, the processes of action selection and inhibition of return, as well as the content and organization of the visual memory. In this sense, it seems better to say that $A_{\mathcal{R}}$ is "specialized" in making only a few, but very good saccades towards an object. These few saccades provide the best estimations that can be achieved. Thus, for this architecture, the more saccades executed and stored in the VMGS, the poorer the quality of estimation on average.

The opposite is the case for $A_{\mathcal{G}}$. This architecture seems to produce poor saccades towards objects in general. Therefore, it requires more saccades in order to get a reasonable sample of gaze space configurations. Thus, the larger this sample the better estimation of the object location.

The difference in accuracy between both architectures with respect to the $\epsilon$-value is hard to derive from the architecture. It is the result of all the processes involved and therefore an emergent property, which for now, we are only able to describe by the statistics derived from our experimental data. Unfortunately, only the same can be said about the error values of visual search when compared with our base-line $2.00 \pm 1.0$ cm [5]. For both architectures, the visual search processes generate higher average errors. These higher error values can only be caused by the visual search processes since the uncertainties of the robotic hardware and the learning errors are indicated by the base-line.

### C. Combining Both Architectures

The fact that both architectures have certain advantages ($A_{\mathcal{G}}$ provides better estimations but needs to execute more saccades, whereas $A_{\mathcal{R}}$ can produce very reasonable estimations with a limited number of saccades) warrants some critical analysis on combining both architectures. Such an integration requires one additional action selection process acting in gaze space, which guarantees conflict resolution if the two streams generate different target coordinates for the eye saccade (Fig. 10).

This extended architecture can also overcome a bottle neck effect that was previously apparent in both architectures $A_{\mathcal{R}}$ and $A_G$; the reach action could only be determined by the content of the visual memory VMGS. Thus, only objects that the system had saccaded to could be picked up. A reach action without a saccade was, therefore, impossible. This limitation is addressed within the combined format by allowing the content of GSSM to be transformed into the reach space. Moreover, like VMGS the GSSM can also be modulated by the reach component. Thus, the system is also able to perform reach guided saccades, i.e., it can look to where the arm reaches.

### *D. Representing Space Through Mappings*

In both architectures, visually guided reaching is achieved through combining two sensorimotor mappings. There is no external absolute coordinate system that the robot system refers to in order to derive spatial locations from the visual data. In former experiments we have shown that these mappings can be learned in a very fast way and are able to adapt continuously to changes, for example, if the spatial location between arm and vision system changes [5], [50]. Therefore, we argue that the introduced architectures establish an embodied representation of space which is generated by the learned sensorimotor mappings and which is entirely the result of robot–environment interaction. Thus, we see this system as a demonstration of how concepts like space can emerge from embodied systems and how, through interaction, it can learn the correlation between different sensorimotor modalities. Here, the different sensorimotor modalities are the active vision system and the robot arm. Both can indeed act independently, but by actively learning the relation between reach and gaze space a new behavioral competence is established: "looking where the arm reaches to," as well as "reaching were the active vision system looks."

### *E. General Framework for Developmental Learning in Humanoid Robots*

At the outset, we promoted the "gaze centered approach" as a step towards developmental learning for humanoid robots. Although the experiments presented here are not conducted on a humanoid robot, we argue that the general framework can directly be applied to any robot system which is equipped with an active vision and arm–hand system. As long as reproducible saccadic eye-movements and reach actions are generated by the system, the gaze centered approach can be implemented. Indeed, a challenge for advanced humanoids is the involvement of head movements when fixating objects and generating eye-saccades. This requires additional mappings and a higher dimensional gaze space. Moreover, the presence of two arm-hand systems needs more sophisticated coordination mechanisms, especially for object manipulation. However, the core concept of the architectures—the gaze centered coordination of vision and reaching remains valid.

The architectures allow the application of any kind of mappings, e.g., neural networks. The action selection processes, here winner-take-all, can be replaced by more sophisticated mechanisms too. In this sense, the two architectures demonstrate the general concept and outline the minimal requirements of our approach in order to achieve the final task, visually guided reaching. As we have already mentioned at the beginning, these particular architectures instantiate a time line similar to the development of these particular competencies in infants, and strongly reflect the interrelationship of primary vision and reach components of the brain. Hence, we argue the architectures provide a robust framework for future studies of developmental learning processes similar to human infants. Linking all the discussed learning processes into one continuum as observed during child development, as opposed to running them in isolated stages, is the focus of current research activities.

## VIII. Conclusion

In this work, we introduced a computational framework that integrates robotic active vision and reaching. The core elements of this framework are three different computational domains (retina, gaze, and reach space), which are linked by sensorimotor mappings. This particular organization is the result of former work in robotics where we have investigated learning schemas for eye saccades and eye-hand coordination inspired by child development and findings in brain research. This framework is, therefore, a first attempt to combine these "pieces" into one framework.

Two architectures were presented that successfully instantiate this framework. The architectures demonstrated how visual search and visually guided reaching can be modulated by the gaze space.

Gaze space modulation allows hand–eye coordination without a global reference frame. The space the robot system interacts with is represented by the two sensorimotor mappings that transform visual data into reach coordinates and vice versa. It could be therefore referred to as an example for a distributed representation of space that emerges from the robot environment interaction and the coordination of different sensorimotor modalities.

We have outlined how this framework leads to a specific time line of competence development when we try to learn the involved sensorimotor mappings from scratch. This time line is similar to child development of humans. Therefore, we argue, this work provides a promising reference architecture for humanoid robotics and developmental learning in humanoids.

## References

[1] 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.

[2] F. Chao, M. H. Lee, and J. J. Lee, "A developmental algorithm for ocular-motor coordination," *Robot. Autonom. Syst.*, vol. 58, pp. 239–248, 2010.

[3] M. Hülse, S. McBride, and M. Lee, "Implementing inhibition of return; embodied visual memory for robotic systems," in *Proc. 9th Int. Conf. Epig. Robot.: Modeling Cogn. Develop. Robot. Syst.*, Venice, Italy, 2009, pp. 213–214.

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

[5] M. Hülse, S. McBride, and M. Lee, "Fast learning mapping schemes for robotic hand-eye coordination," *Cogn. Comput.*, vol. 2, no. 1, pp. 1–16, 2010.

[6] J. Gottlieb, P. Balan, J. Oristaglio, and M. Suzuki, "Parietal control of attentional guidance: The significance of sensory, motivational and motor factors," *Neurobiol. Learn. Memory*, vol. 91, no. 2, pp. 121–128, 2009.

[7] M. Hülse, S. McBride, and M. Lee, "Gaze modulated visual search for active vision," in *Proc. 11th Towards Autonom. Robot. Syst. (TAROS 2010)*, Plymouth, U.K., 2010, pp. 83–90.

[8] M. Hülse and M. Lee, "Adaptation of coupled sensorimotor mappings: An investigation towards developmental learning of humanoids," in *Proc. 11th Int. Conf. Simul. Adapt. Behav. (SAB)*, Paris, France, 2010, LNAI 6226, pp. 468–477.

[9] B. Stein and M. Meredith, *Functional Organization of the Superior Colliculus*. Hampshire, U.K.: Macmillan, 1991, pp. 85–100.

[10] C. M. Harris, M. Jacobs, F. Shawkat, and D. Taylor, "The development of saccadic accuracy in the first seven months," *Clinical Vis. Sci.*, vol. 8, no. 1, pp. 85–96, 1993.

[11] V. Casagrande, I. Diamond, J. Harting, W. Hall, and G. Martin, "Superior colliculus of the tree shrew- structural and functional subdivision into superficial and deep layers," *Science*, vol. 177, no. 4047, pp. 444–447, 1972.

[12] J. Gottlieb, "From thought to action: The parietal cortex as a bridge between perception, action, and cognition," *Neuron*, vol. 53, no. 1, pp. 9–16, 2007.

[13] B. Stein, M. Wallace, T. Stanford, and W. Jiang, "Cortex governs multisensory integration in the midbrain," *Neuroscientist*, vol. 8, no. 4, pp. 306–314, 2002.

[14] J. McHaffie, T. Stanford, B. Stein, W. Coizet, and P. Redgrave, "Subcortical loops through the basal ganglia," *Trends Neurosci.*, vol. 28, no. 8, pp. 401–407, 2005.

[15] H. Cui and R. Andersen, "Posterior parietal cortex encodes autonomously selected motor plans," *Neuron*, vol. 56, pp. 552–559, 2007.

[16] C. Buneo, M. Jarvis, A. Batista, and R. Andersen, "Direct visuomotor transformations for reaching," *Nature*, vol. 416, no. 6881, pp. 632–636, 2002.

[17] R. Clifton, D. Muir, D. Ashmead, and M. Clarkson, "Is visually guided reaching in early infancy a myth," *Child Develop.*, vol. 64, no. 4, pp. 1099–1110, 1993.

[18] M. Johnson, D. Mareschal, and G. Csibra, *The Development and Integration of Dorsal and Ventral Visual Pathways in Object Processing*. Cambridge, MA: MIT Press, 2008, ch. 28, pp. 467–478.

[19] J. Bruner, *Acts of Meaning*. Cambridge, MA: Harvard Univ. Press, 1990.

[20] J. Rutkowska, "Scaling up sensorimotor systems: Constraints from human infancy," *Adapt. Behav.*, vol. 2, pp. 349–373, 1994.

[21] E. Thellen, "Motor development," *Amer. Psychol.*, vol. 50, pp. 79–95, 1995.

[22] G. Butterworth and M. Harris, *Principles of Developmental Psychology*. Hove, U.K.: Erlbaum, 1994.

[23] R. J. Adams, M. L. Courage, and M. E. Mercer, "Systematic measurement of human neonatal color vision," *Vis. Res.*, vol. 34, pp. 1691–1701, 1994.

[24] P. J. Kellman and M. E. Arterberry, *The Cradle of Knowledge: Development of Perception in Infancy*. Cambridge, MA: MIT Press, 1998.

[25] J. Oates, C. Wood, and A. Grayson, *Psychological Development and Early Childhood*. Malden, MA: Blackwell, 2005.

[26] M. D. Sheridan, *From Birth to Five Years: Childern's Developmental Progress*. Windsor, U.K.: NFER-NELSON, 1973.

[27] D. Muir and J. Field, "Newborn infants orient to sounds," *Child Develop.*, vol. 50, pp. 431–436, 1979.

[28] J. F. Rosenblith, *In the Beginning: Development from Conception to Age Two*, 2nd ed. Newbury Park, CA: Sage, 1992.

[29] R. N. Aslin, "Visual and auditory development in infancy," in *Handbook of Infancy*, J. D. Osofsky, Ed., 2nd ed. New York: Wiley, 1987.

[30] E. Tronick, "Stimulus control and the growth of the infant's effective visual field," *Percept. & Psychophys.*, vol. 11, no. 5, pp. 373–376, 1972.

[31] D. Maurer and C. Maurer, *The World of the Newborn*. New York: Basic Books, 1988.

[32] E. Tronick and C. Clanton, "Infant looking patterns," *Vis. Res.*, vol. 11, pp. 1479–1486, 1971.

[33] P. Salapatek, "Pattern perception in early infancy," in *Infant Perception from Sensation to Cognition*, L. B. Cohen and P. Salapatek, Eds. New York: Academic, 1975, vol. 1, Basic Visual Processes.

[34] M. M. Haith and G. S. Goodman, "Eye movement control in newborns in darkness and in unstructured light," *Child Develop.*, vol. 53, pp. 974–977, 1982.

[35] A. E. Milewski, "Infants' discrimination of internal and external pattern elements," *Exp. Child Psychol.*, vol. 22, pp. 229–246, 1976.

[36] M. R. Fiorentino, *A Basis for Sensorimotor Development, Normal and Abnormal: The Influence of Primitive, Postural Reflexes on the Development and Distribution of Tone*. Springfield, IL: Charles C. Thomas, 1981.

[37] F. Goodkin, "The development of mature patterns of head–eye coordination in the human infant," *Early Human Develop.*, vol. 4, pp. 373–386, 1980.

[38] J. I. P. De Vries, G. H. A. Visser, and H. F. R. Prechtl, "Fetal motility in the first half of pregnancy," in *Continuity of Neural Function from Prenatal to Postnatal Life*, H. F. R. Prechtl, Ed. London, U.K.: Spastics, 1984.

[39] J. Piaget, *The Origin of Intelligence in the Child*. Evanston, IL: Routledge, 1952.

[40] B. Wyke, *The Neurological Basis of Movement – A Developmental Review*, K. S. Holt, Ed. Philadelphia, PA: Lippincott, 1975.

[41] G. Butterworth and B. Hopkins, "Hand-mouth coordination in the newborn baby," *Brit. J. Develop. Psychol.*, vol. 6, pp. 303–314, 1988.

[42] P. Rochat, E. M. Blass, and L. B. Hoffmeyer, "Oropharyngeal control of hand-mouth coordination in newborn infants," *Develop. Psychol.*, vol. 24, pp. 459–463, 1988.

[43] C. Trevarten, "The psychobiology of speech development," in *Language and Brain: Developmental Aspects [Special Issue]*, E. Lennenberg, Ed. Cambridge, MA: MIT Press, 1974, Neurosciences Research Program Bulletin, pp. 570–585.

[44] C. von Hofsten, "Eye-hand coordination in newborns," *Develop. Psychol.*, vol. 18, pp. 450–461, 1982.

[45] C. von Hofsten, "Developmental changes in the organisation of prereaching movements," *Develop. Psychol.*, vol. 20, pp. 378–388, 1984.

[46] M. M. Shirley, *The First Two Years – A Study of Twenty-five Babies*. Minneapolis: Univ. Minnesota Press, 1933, vol. 2, Intellectual Development.

[47] E. Thelen, D. Corbetta, K. Kamm, J. P. Spencer, K. Schneider, and R. F. Zernicke, "The transition to reaching: Mapping intention and intrinsic dynamics," *Child Develop.*, vol. 64, no. 4, pp. 1058–1098, 1993.

[48] B. L. White, P. Castle, and R. Held, "Observations on the development of visually-directed reaching," *Child Develop.*, vol. 35, no. 2, pp. 349–364, 1964.

[49] H. Hoffmann, W. Schenk, and R. Möller, "Learning visuomotor transformations for gaze-control and grasping," *Biol. Cybern.*, vol. 93, pp. 119–130, 2005.

[50] M. Lee, Q. Meng, and F. Chao, "Developmental learning for autonomous robots," *Robot. Autonom. Syst.*, vol. 55, no. 9, pp. 750–759, 2007.

**Martin Hülse** (M'09) received the Dipl.Inf. degree in computer science from the Friedrich-Schiller-University, Jena, Germany, in 2000, and the doctoral degree (Dr. rer. nat.) from the University of Osnabrück, Germany, in 2006.

He is currently member of the Intelligent Robotics Group at Aberystwyth University, Wales, U.K. From 2002 to 2006, he worked at the Fraunhofer Institute for Autonomous Intelligent Systems (now FhI-IAIS). He has been engaged in national and international robotic related research and development projects since 2000. His main research interests are embodied cognition, complex adaptive systems, neural computation, and machine learning and their applications in the field of adaptive behavior and autonomous robots.

**Sebastian McBride** received the B.Sc. degree in zoology from the University of Liverpool, Liverpool, U.K., in 1992, and the Ph.D. degree in animal behavior from the Royal School of Veterinary Studies, University of Edinburgh, Edinburgh, U.K., in 1999.

He has been lecturing and researching in the area of behavioral neurophysiology since 1993 and has recently joined the Intelligent Robotics Group at Aberystwyth University to develop architectures based on the current knowledge of brain systems.

**James Law** (S'04–M'07) received the M.Eng. degree from the University of Hull, Yorkshire, U.K., in 2003, and the Ph.D. degree from the Open University, Milton Keynes, U.K., in 2008. He is currently a Postdoctoral Research Associate in the Intelligent Robotics Group at Aberystwyth University, Wales, U.K.

He is currently researching developmental models as part of the European Framework 7 IM-CLeVeR project. His research interests include biologically and psychologically constrained architectures for robotic learning and control, and complex and adaptive intelligent systems.

Dr. Law has been a member of the IEEE Robotics and Automation Society since 2004.

**Mark Lee** received the B.Sc. and M.Sc. degrees in electrical engineering from the University of Wales, Swansea, U.K., in 1967 and 1969, respectively. He received the Ph.D. degree in psychology from Nottingham University, Nottingham, U.K., in 1980.

He is currently a Professor of Intelligent Systems in the Department of Computer Science at Aberystwyth University, Wales, U.K. He is also currently a Principal Investigator on several EPSRC and EU funded research projects on robotic sensory-motor learning, adaptation, and development. His main research interests are in developmental robotics, inspired by early infant psychology.