An evaluation of gaze modulated spatial visual search for robotic active vision

Hülse, Martin; McBride, Sebastian; Lee, Mark

Published in:
TAROS 2010

Publication date:
2010

Citation for published version (APA):
Hülse, M., McBride, S., & Lee, M. (2010). An evaluation of gaze modulated spatial visual search for robotic active vision. In *TAROS 2010: Proceedings of Towards Autonomous Robotic Systems 2010* (pp. 83 - 90). Plymouth University Press. http://hdl.handle.net/2160/5699

An evaluation of gaze modulated spatial visual search for robotic active vision

Martin Hülse, Sebastian McBride and Mark Lee

Abstract— **Active vision is an essential part of many autonomous robot systems, in particular humanoid robots. In this work we present a method for spatial visual search which is modulated by the absolute motor positions of the active vision system resulting from saccades to objects. A central element of this approach is the so called *visual memory* where these motor configurations are stored. Based on these motor data, the system can evaluate which of the current visual stimuli have already been saccaded to. In this sense, motor configurations in the visual memory modulate the selection of visual targets for the purpose of saccade. Two architectures are presented which instantiate this gaze modulated visual search in a robotic scenario. The paper also presents a series of systematic experiments demonstrating the impact of two essential parameters (ϵ [size of inhibitory neighborhood] and γ [decay rate]) on the behavioural dynamics of the active vision system. ϵ was found to determine the number of saccades needed to scan a scenario whilst γ controlled the persistence of visual memory. Finally, we discuss the advantage of gaze modulated visual search compared to other common strategies without gaze-modulation. It is apparent that gaze space modulation is advantageous with respect to real-time performance and scalability, and therefore offers an interesting alternative approach for active vision in robotics as well as for general models of visual search.**

I. Introduction

Visual perception in humans and most biological systems is an active process [2], [3], [7], [9]. Summarized by the term *Animate Vision* the computational advantages of having the ability to "control the direction of gaze" [2] were already illustrated almost 20 years ago. Even without highlighting the behavioural aspects of Animate Vision and their fundamental implications for Artificial Intelligence and Cognitive Science, from a pure engineering perspective, active vision is a valuable approach because it provides a robot system with a nearly unlimited field of view. Moreover, for a meaningful interaction of humans and robots it is important to have a direct action understanding. A robotic active vision system supports action understanding by indicating the system's focus of attention when fixating an object or getting attracted by specific physical stimuli. In this sense through an active vision system the robot state becomes intuitively readable by humans [4].

One of the challenges when dealing with robotic active vision systems is to go beyond the purely reactive nature of gaze control. This is crucial for visual search tasks as we understand them: a systematic fixation of objects in the environment without getting caught by one object. This can easily happen when similar objects are present and the local image data can't be related to a global reference frame. In the majority of models on human vision this problem is ignored by assuming or creating a global reference. Experiments of visual search tasks are very often conducted in a setup where a global retinotopic reference frame is implicitly or explicitly defined either by static images (e.g., [7]) or by the exploitation of the specifics of computer simulations, e.g. [10], [1].

An active vision system in an unconstrained environment doesn't have direct access to a global retinotopic reference frame. Our solution to this problem is to make use of the configuration of the gaze control system. Thus, we call this approach *gaze-modulated spatial visual search*. This approach is best motivated by a brief introduction of the common computational models for visual search.

A. Computational models for visual search

Many computational models of human visual search assume separated processes for object location and object identification. This is in line with findings from Brain Research on the organisation of the visual cortex in ventral ("what") and dorsal ("where") pathways. According to the model proposed by Ballard and colleagues [3] visual object identification in humans is done only when the object is in the fovea which is the central part of the eye providing the highest resolution. An eye movement which brings an object into the fovea (fixation) is also called a saccadic eye movement or saccade. According to Ballard a comprehensive visual search which includes complex object features and object location can only be provided by a serial search [3]. Or, in other words, visual search involves the successive fixation of the objects in the environment whilst an inhibition of return mechanism (IOR) guarantees that saccades are not repeatedly executed towards the same object.

Recent studies [9] have shown that saccades in humans are modulated by the task and the context. This might be seen as contradicting other popular models of visual search highlighting bottom up saliency maps [7]. Saliency maps are exclusively generated in a bottom up manner where specific combinations of low-level image features determine the rank of image regions according to their attractiveness or saliency in the given image. Alternative models are proposed which are able to combine top-down and bottom up approaches by modulating the saliency map generation processes in a top-down manner [5]. In such a way, saliency maps reflect a biased view of the visual data towards pre-defined feature combinations.

Dept. Computer Science, Aberystwyth University, Penglais, SY23 3DB, Wales, UK [msh,sdm,mhl]@aber.ac.uk

No matter how top-down and bottom-up approaches are integrated in a comprehensive and biological plausible way, for a robotic active vision system it is important to emphasize that the active nature of serial visual search alters inevitably the visual input. This is because a saccade of an active vision system moves the camera and this changes significantly the visual input. Unfortunately, alterations of visual input data due to camera or eye movements aren't considered in most of the literature on bottom-up approaches towards visual search. The proposed IOR mechanisms usually assume static image data where image regions in the sense of X-Y-coordinates can easily be labeled as regions which already have been a target of an eye-saccade [7].

Such mechanisms, however, must fail for active vision systems because image regions represented as X-Y-coordinates before the camera movement doesn't refer to same objects in the environment after the camera movement. In addition, most bottom-up approaches assume not only static image data but also a global retinotopic reference frame as the action selection domain for eye-saccades. Here, local image data is matched against the global retinotopic reference frame.

Trying to follow such an approach for a robotic active vision system, where only data in a local retinotopic reference frame can be accessed at any one time, there is the need to match local retinotopic data into a global retinotopic reference frame in order to solve the problem of comparing this data against previously saccaded to objects for the purposes of IOR. Such processes involve high computational costs (memory and time) and sophisticated calibration processes while having weak robustness and real-time performance.

This issue has previously been addressed by Alexandre and colleagues [1], [11]; here, before a saccade is executed, the expected change of the local image data is anticipated in order to evaluate which of the salient image regions the vision system has already saccaded to. This approach, however, has two disadvantages. Firstly, it doesn't meet the real-time constraints needed for highly interactive robotic systems acting in changing environments [11]. Secondly, it operates on a local retinotopic reference frame only.

The real-time constraints aren't met because the computational costs of the anticipation process operating on the whole image data, i.e., the higher the resolution the higher the computational costs, and consequently the slower the robot system. The second problem is caused by the local domain of the anticipation process as it is only operating on the current image data. Thus, IOR works only for the local retinotopic reference frame and objects outside the current field of view can't be processed. The consequence of this is that, as soon as a previously saccaded to object disappears from the field of view due to camera movement, it is "forgotten". Returning the camera to a similar position then results in re-saccade with the object being treated as new.

B. *Specific aims*

The brief discussion of current computational models for visual search under consideration of robotic active vision systems clearly illustrates two requirements for a successfully IOR mechanism. Firstly, the IOR domain must operate in a global reference frame in order to keep track of objects not present in the current local image data. Secondly, an IOR domain shouldn't be the global retinotopic reference frame since it isn't directly available to the active vision system. Furthermore, explicit generation of such a model should be avoided due to the computational costs, level of robustness and real-time constraints associated with this approach.

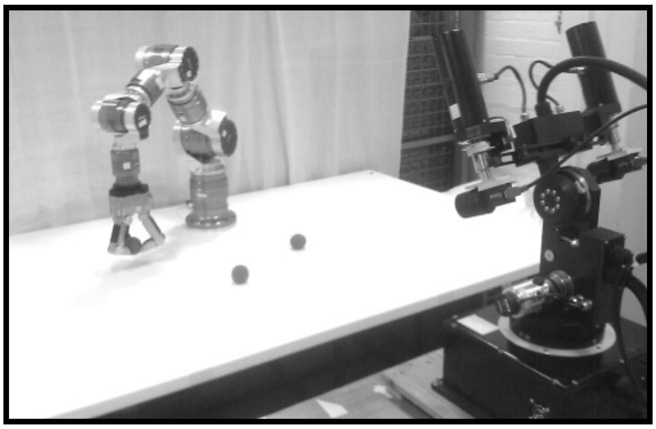

Fig. 1. Robotic scenario, where an active vision system observes the objects on the table, which can be pick-up by a manipulator.

The objective of this paper is, therefore, the demonstration of inhibition of return processes which are modulated by the motor configuration of the active vision system, which we refer to here as the gaze space. Hence, visual search emerges from the interaction between local retinotopic image data and the gaze space configurations of successful saccades. We introduce two computational architectures that demonstrate that the gaze space provides the required global reference frame for visual search while meeting the real-time constraints that guarantee high robot-environment interaction dynamics.

II. Methods

The active vision system consists of two cameras (both provide RGB 1032x778 image data) mounted on a motorised pan-tilt-verge unit (Figure 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). Thus, the active vision system configuration is fully determined by the absolute motor positions of the tilt and left verge axis, (p_{tilt}, p_{vL}). The absolute positions of these two parameters define the *gaze space*.

A. *Two computational architectures for gaze modulation*

In the following, we introduce two computational architectures for gaze-modulated visual search. Both architectures use a mapping process to facilitate saccade action where X-Y-coordinates of the local retinotopic image data are transformed into motor position changes $(\Delta p_{tilt}, \Delta p_{vL})$, given in

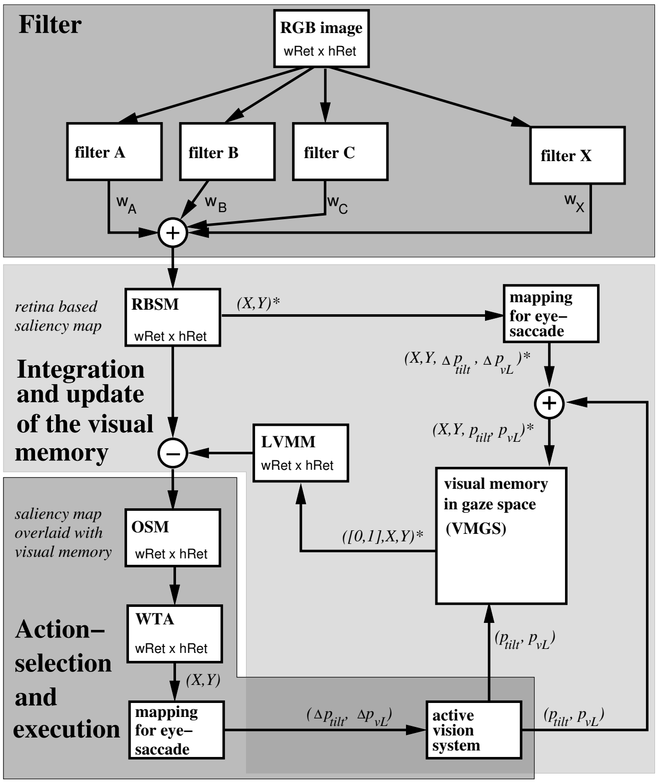

Fig. 2. Computational architecture $A_{\mathcal{R}}$

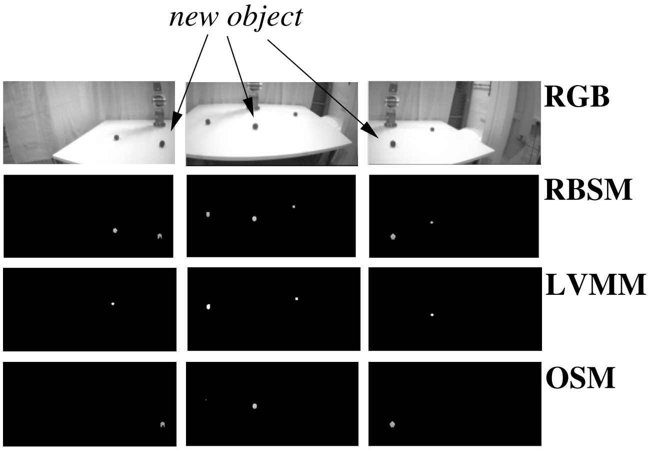

Fig. 3. Particular system states of architecture $A_{\mathcal{R}}$ for different camera positions after a new object is placed on the table. The new object is not yet present in the visual memory therefore it is the only stimuli in the OSM. The stimuli representing the old objects are present in LVMM which inhibits their emergence OSM. See text for details.

rad. The execution of these motor position changes drive the camera in such a way that the corresponding stimulus at the X-Y-coordinates end up in the fovea, i.e. the image centre. The actual saccade mappings can either be learned [6], [8] 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_{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, referred to as VMGS (visual memory in gaze space).

Obviously, the visual search is modulated by the content of the visual memory (VMGS). We now introduce two very distinct strategies how this modulation can generate an IOR mechanism. In the first architecture $A_{\mathcal{R}}$ Figure 2, 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 $A_{\mathcal{G}}$ Figure 4, 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 global gaze-space for $A_{\mathcal{G}}$. The following sections provide a more detailed descriptions of the two architectures.

B. 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 colour filtering process of the current camera image data generates a saliency map referred to as the retina-based saliency map (RBSM). The dimension of RBSM is determined by the width and height of the camera images (wRet x hRet). Each stimuli in RBSM is represented by a non-zero entry of the corresponding X-Y-coordinates. Due to the previously learnt eye-saccade mapping, each X-Y-coordinate of a non-zero entry in the RBSM derives a corresponding motor changes $(\Delta p_{tilt}, \Delta p_{vL})$ for a successful saccade. Together with current absolute motor positions (delivered by the active vision system), this produces, for each non-zero X-Y-coordinates, the expected absolute motor positions of the vision system if a saccade towards this stimulus was executed. This is expressed in Fig. 2 in the form of $(X, Y, p_{tilt}, p_{vL})^*$ which refers to a list of all non-zero X-Y-coordinates and their expected final absolute motor configuration 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 with 1, otherwise 0. In this way we get a new list $([0/1], X, Y)^*$ of all non-zero X-Y-coordinates in the RBSM which are labeled according to their presence in the VMGS. This list can be transformed into a map LVMM (local visual memory map) having the same dimensions as RBSM. In contrast to RBSM, the non-zero entries in LVMM represent the stimuli the system has already saccaded to. Hence, the subtraction of RBSM by LVMM will generate a new map OSM (overlaid saliency map) which contains only the stimuli which haven't yet been saccaded to by the active vision system.

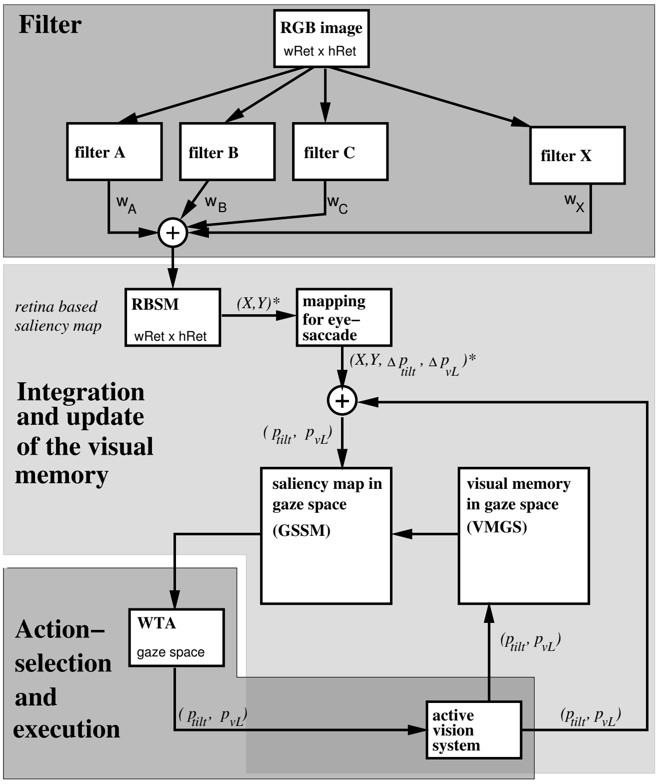

Fig. 4. Computational architecture $A_{\mathcal{G}}$

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_{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. Figure 3 represents the image data (RGB) as well as the RBSM-, LVMM- and OSM-data for different camera positions. The non-black entries represent stimuli or non-zero 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 (via 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 Figure 3), 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.

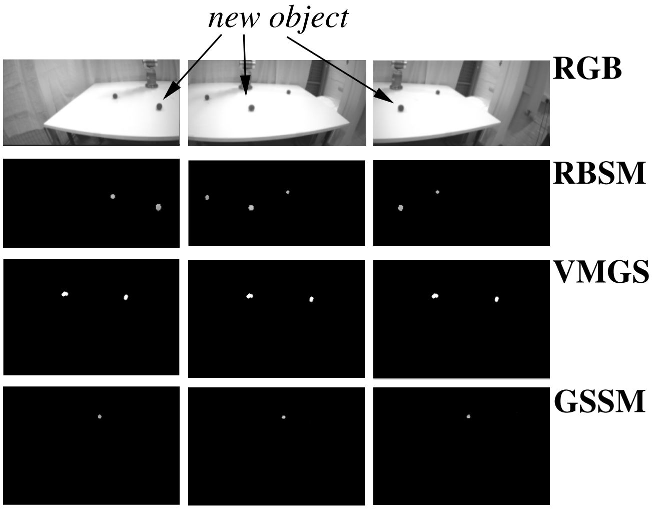

Fig. 5. Particular system states of architecture $A_{\mathcal{G}}$ for different camera positions after a new object is placed on the table. The new object is not yet present in the visual memory VMGS. Since VMGS directly inhibits the gaze space based saliency map, only one stimulus is present in GSSM. See text for details.

C. Action selection in gaze space

The second architecture $A_{\mathcal{G}}$ has a structure similar to $A_{\mathcal{R}}$, Figure 4 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 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_{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 position (p_{tilt}, p_{vL}). If the movement of the camera into 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 have the same domain. Thus, inhibition of return 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 (Figure 5). Like in the scenario above, we started with two objects on the table. After the two object 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.

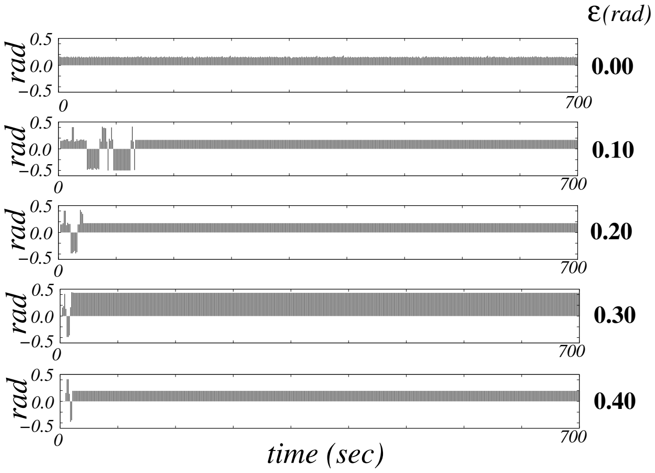

Fig. 6. Test of architecture $A_{\mathcal{R}}$ for different ϵ-values while the decay mechanism was deactivated. (Data shown for only 700 seconds.)

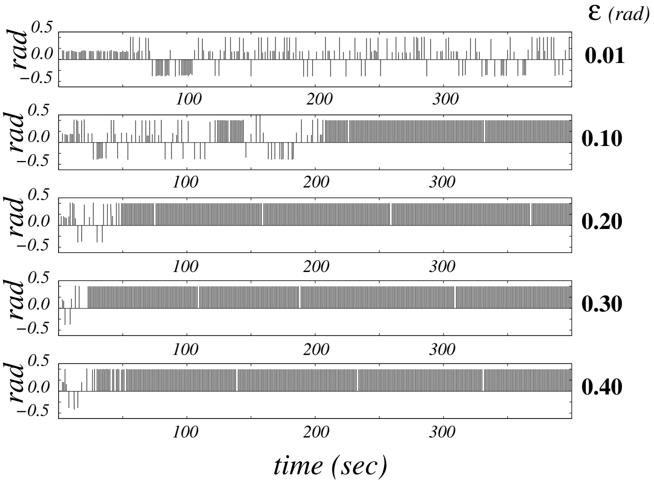

Fig. 7. Test of architecture $A_{\mathcal{G}}$ for different non-zero ϵ-values while the decay was deactivated. (Data shown for only 400 seconds.)

D. Experimental conditions

In implementing the two architectures $A_{\mathcal{R}}$ and $A_{\mathcal{G}}$ for our robotic system, we have to consider two essential parameters. The first parameter (ϵ), 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. ϵ, therefore, defines a neighbourhood for a specific gaze space configuration to compensate for this noise effect.

The second essential parameter γ determines the decay of the entries in the visual memory VMGS. Once a configuration is stored in the visual memory, the action selection process for eye-saccade is "blind" to this spatial location. Hence, in order to keep the robotic system up-to-date in a changing environment we introduce a decay variable. Each entry in the visual memory has an "age" value a and this value is increased by 1 in each iteration step with the average update rate set at 40 Hz. The parameter γ defines the maximum age value. If this value is reached the entry is removed from the visual memory. Furthermore, the older the entry the less the corresponding activation in visual memory VMGS, and therefore the less the inhibition of OSM (in $A_{\mathcal{R}}$) and GSSM (in $A_{\mathcal{G}}$), respectively.

The relation between current age value a and activation value in VMGS is linear. Each new entry in the visual memory has an age value of zero which corresponds to maximal activation value, i.e. maximal suppression. The maximum age γ results in zero activations, i.e. non suppression, since these values are removed from the VMGS.

Obviously, γ and ϵ will highly influence the behavioral dynamics of the active vision system. The next section describes the experimental results for different parameter configurations for both architectures.

III. Experimental Results

Two series of experiments for each architecture were conducted to test different values of one parameter while the other was fixed. In each experiment three object were placed on the table, similar to the scenario shown in Fig. 3 and 5. For each parameter setting, the visual memory VMGS contained zero values and the starting orientation of the camera was kept constant.

Measuring saccades was done via output recording of the absolute positions of the verge motor. These verge values are distinct for each object when saccaded to (≈ -0.5, ≈ 0.2, and ≈ 0.5 rad). Unfortunately, the diagrams in Fig. 6, 7, 8, and 9 can only provide a qualitative picture of the system behavior. Whether or not the camera is moving can not be derived from these data set. However, in the first two Figures (Fig. 6, 7) for $\gamma > 0$ one can see the substantial difference in time it needs before the system comes to an halt indicating complete spatial inhibition of all visual stimuli. While in Fig. 8 and 9 the recorded verge positions provide a fair impression how frequently the system executes a saccade over a given period of time.

Due to process overload of the running models and the lower priority of the recording process, some gaps in the data were (as illustrated in the plots) were occasionally observed.

A. Different ϵ-value and no decay

In this set of experiments different ϵ values were tested while the decay process of the visual memory VMGS was deactivated. Hence, once a (p_{tilt}, p_{vL})-configuration was stored in the VMGS, it remained there. Thus, the visual search will saccade to all the stimuli until the stored configurations in VMGS cover the whole scene, at which point the system will stop. The metric of the ϵ-parameter was the Euclidean distance and each test run was for over 800 seconds.

All data are presented in Figures 6 and 7. The first experiment set $\epsilon = 0.0$ and essentially illustrates system behaviour without an IOR 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 from

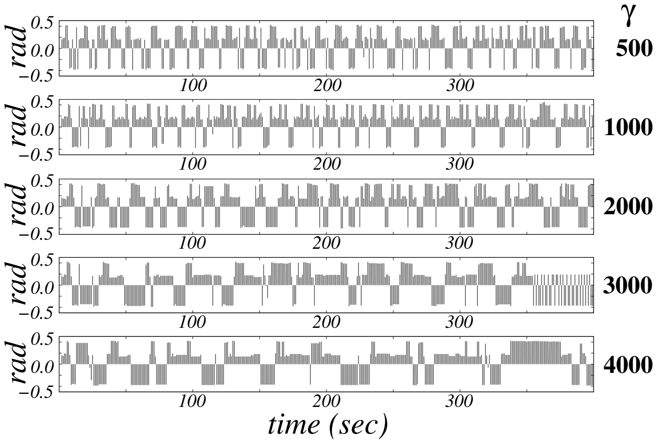

Fig. 8. Test of architecture $A_{\mathcal{R}}$ for different decay values γ for $\epsilon = 0.25$.

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 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 ϵ-value the fewer saccades or (p_{tilt}, p_{vL})-configurations in VMGS were necessary to inhibit the stimuli generated by the objects. This was indicated by the time and the number of saccades required until the active vision system stopped. Although a qualitatively similar trend of behaviour was generated by architecture $A_{\mathcal{G}}$ (Fig. 7), the time taken to reach execution of the final saccade was significantly longer ($p < 0.05$) compared to architecture $A_{\mathcal{R}}$. (See Table I for numerical values.)

B. Different decay values γ for fixed ϵ-values

In this series of experiments different decay values γ were tested for $\epsilon = 0.25$ with each experiment run over 400 seconds. In contrast to the experiment above, because of the presence of a decay, the system continuously saccaded to the objects on the table even after all objects had previously been saccaded to.

For both architectures, it was apparent that increasing the γ value resulted in fewer saccades. For example, referring to Fig. 8, $\gamma = 4000$ resulted in the camera remaining in the same position for a much longer period of time (i.e. there were less saccades) compared to $\gamma = 500$. This was also observed for architecture $A_{\mathcal{G}}$, see Figure 9. Further generation of data will facilitate a more detailed quantitative comparison between the two architectures.

ϵ	time (sec.) until final saccade	
	$A_{\mathcal{R}}$	$A_{\mathcal{G}}$
0.10	103	213
0.20	33	50
0.30	26	23
0.40	10	54

TABLE I

DURATION OF SACCADING PROCESS WITHOUT DECAY

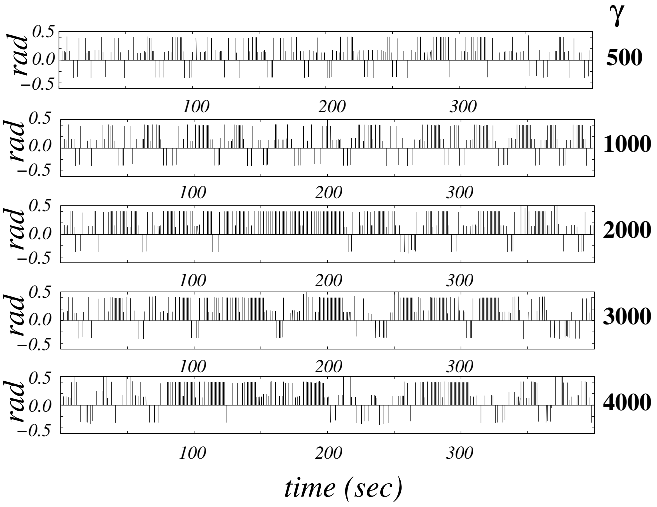

Fig. 9. Test of architecture $A_{\mathcal{G}}$ for different decay values γ for $\epsilon = 0.25$.

IV. DISCUSSION

A. Visual search

The experiments have demonstrated that visual search of a robotic active vision system can be modulated by the gaze space. The gaze space provides a global reference frame which is essential for an active vision system which only has access to a local (retinotopic) reference frame. Our experiments have shown the impact of the two parameters γ and ϵ on the behaviour of the active vision system in a static scenario. Preliminary statistical analysis confirmed this prediction for the ϵ parameter but further work needs to be done with regard to the γ parameter from a quantitative perspective.

B. Computational costs

In both architectures, visual data are mapped between retina and gaze space. These mappings are solely for the stimuli (non-zero entries) in RBSM and not for the whole image data (all pixels). This creates an essential computational advantage since the the number of non-zero entries are generally much less compared to the total number for the entire image. Hence, our method scales better with respect to high resolution image data compared to other work [1], [11].

The processing of the visual memory VMGS also needs very low computational resources because only the successful saccades are stored. Even without a decay mechanism, this number is substantially smaller than the data needed to build up a global retinotopic reference frame of reasonable quality. Hence, for very little computational costs the gaze space modulation provides a global reference frame for the visual search.

C. Robustness

All the entries in the visual memory represent valid spatial object locations. This is guaranteed by testing the success of a saccade explicitly, i.e. after a saccade the RBMS must have non-zero activity in a pre-defined central region of the image. Such a test is necessary for both architectures since the more stimuli that are inhibited by the visual memory, the more sensitive the system becomes to noise. Within both architectures, noise-initiated configurations can easily be eliminated to keep the visual memory clear of these artefacts through this aforementioned test.

D. Context sensitive ϵ-values

The experiments have shown that the larger the ϵ value the fewer saccades are required towards the same object in order to cover the corresponding area of stimuli. However, if the ϵ value is too high and the IOR inhibits too large an area, the system become "blind" to specific objects. Therefore, it is important that a reasonable ϵ-value is chosen with respect to the specifics of the given scenario. This includes the object sizes, the working area where objects are located and the required level of precision.

However, a problem also exists with having a constant ϵ-values in that the size of the non-zero activation area in the RBSM is not only determined by the physical size of the object but also by the distance between the object and camera. Thus, although objects on the table have the same size, a different number of saccades might be necessary for each depending on the distance. For more complex scenarios, a more variable ϵ-value might therefore be recommended with the actual object size as well as the distance between camera and object determining the ϵ-value. The former could be derived from higher level object feature detection processes and the latter hard-wired with respect to the actual region in the gaze space.

E. Dealing with moving objects

For non-static cameras it is particularly important to provide mechanisms which allow a robust detection of moving objects. The challenge is to distinguish between image changes caused by "ego-motion" of the camera and changes that occur due to moving objects. Architecture $A_{\mathcal{G}}$ already provides a solution to this problem since the stimuli in RBSM are mapped immediately to a global reference frame (GSSM). Thus, if the camera moves, the static object doesn't change its position in this reference frame. However, if the stimulus changes position within the GSSM, then this must be due to external motion of that object and not by the moving camera. In other words, the transformation of the visual input into the gaze space eliminates the ego-motion of the camera which makes the detection of moving object a straight forward procedure.

F. Capacity of the visual memory

The introduced decay mechanisms obviously limits the system's capacity with respect to the number of stimuli that can be stored in the visual memory. This will obviously limit the number of objects the system can store in the visual memory because only additional saccades towards the already stored objects can compensate the decay of formerly stored configurations. In general we can say that large γ- or ϵ-values will increase the capacity of the visual memory for stored objects.

G. Interplay of local and global domains

Finally, it is interesting to compare our two architectures $A_{\mathcal{R}}$ and $A_{\mathcal{G}}$ for gaze modulated visual search with the two non-gaze modulated approaches which we described briefly in the Introduction section, the bottom-up approach first introduced by Itty and Koch [7] and the work on visual search in robotics by Alexandre and colleagues [1], [11]. Most enlightening is such a comparison with respect to the domains of the different processing tasks involved in the visual search, namely (1) action selection, (2) of the visual memory, and (3) the domain the IOR is operating on, i.e. where the suppression of stimuli takes places.

For pure bottom-up approaches we saw that action selection, i.e. selection of the image region to be saccaded to, takes place in a global retinotopic reference frame. Consequently this reference frame is also used to label regions already saccaded to and to suppress them in order to allow a saccading to the next salient image region. Hence, the complete information process of visual search is represented in a global retinotopic reference frame. Almost the same can be said about the approach of Alexandre and colleagues, but here it is the local retinotopic reference frame where all the processes of the visual search are represented. The two architectures presented within this study also make use of a global reference frame, the gaze space, in order to represent the visual memory. However, for architecture $A_{\mathcal{R}}$, action selection is done at the level of the local retinotopic domain (OSM in Fig. 2) 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 because the stimuli in GSSM are determined by the stimuli in RBSM. The suppression of GSSM by VMGS, however, acts in the global domain of the gaze space. Table II provides a summary of these approaches.

The advantage of the architectures introduced in this paper is the interplay between local and global reference frames. This is possible because of the mapping between the local retinotopic and the global gaze space. As we have seen in architecture $A_{\mathcal{R}}$, this mapping can even be bidirectional,

approach to		domain of	
visual search	action selection	visual memory	IOR
bottom up	global RT	global RT	global RT
Alexandre et al.	local RT	local RT	local RT
$A_{\mathcal{R}}$	local RT	global GS	local RT
$A_{\mathcal{G}}$	local GS	global GS	global GS

TABLE II

OVERVIEW OF THE DISCUSSED APPROACHES OF VISUAL SEARCH

(RT... retinotopic reference frame, GS... gaze space reference frame) from RBSM (retina space) to VMGS (gaze space) to LVMM and OSM (both retina space).

V. Conclusion

We have demonstrated that spatial visual search for active robotic vision can be modulated by the gaze space. The central element of this process is the visual memory whereby it stores the motor configuration of the active vision system resulting from saccadic eye-movements. Based on these data, the system can evaluate which of the current visual stimuli it has already saccaded to. In this sense the visual memory provides a global reference frame for the active vision system although only a local retinotopic reference frame is directly accessible.

Two architectures for gaze space modulated visual search were introduced and their performance presented in a series of experiments. Both architectures make use of the visual memory but differ with respect to the action selection domain of the saccadic eye-movement: local retinotopic reference frame (architecture $A_{\mathcal{R}}$) and local gaze space reference frame (architecture $A_{\mathcal{G}}$).

We have outlined that the visual memory in gaze space requires less computational resources (memory and time) than similar processing tasks in a global retinotopic reference frame. Therefore we argue, gaze space modulation performs much better with respect to real-time constraints and scalability.

It was shown that the behavioural dynamics of the visual search is essentially determined by two parameters ϵ and γ. The number of saccades needed to scan a scenario is determined by ϵ whilst γ controls the persistence of visual memory.

In summary, this work highlights gaze modulation for active vision systems as a promising alternative for spatial visual search in robotics. It might also offer plausible mechanisms for more general models of visual search with future developments integrating object identification through the processing and representation of more complex object features. Such a model would allow visual search to be biased not only by spatial location but also by specific object features.

VI. ACKNOWLEDGMENTS

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

References

[1] F. Alexandre. Cortical basis of communication: Local computation, coordination, attention. *Neural Networks*, 22, 126-133, 2009.

[2] D.H. Ballard. Animate vision. *Artificial Intelligence*, 48:57–86, 1991.

[3] D.H. Ballard, M.M. Hayhoe, P.K. Polly, and R.P.N. Rao. Deictic codes for the embodiment of cognition. *Behavioral and brain sciences*, 20:723–767, 1997.

[4] C. Breazeal, A. Edsinger, P. Fitzpatrick, and B. Scassellati. Active vision for sociable robots. *IEEE TRANSACTIONS ON SYSTEMS, MAN, CYBERNETICS ANDPART A: SYSTEMS AND HUMANS*, 31:443–453, 2001.

[5] S. Frintrop. *VOCUS: A Visual Attention System for Object Detection and Goal-Directed Search*. Springer, 2006.

[6] H. Hoffmann, W. Schenk, and R. Möller. Learning visuomotor transformations for gaze-control and grasping. *Biol Cybern*, 2005.

[7] L. Itti and Ch. Koch. A saliency-based search mechanism for overt and covert shifts of visual attention. *Vision Research*, 40:1489–1506, 1999.

[8] M.H. Lee, Q. Meng, and F. Chao. Developmental learning for autonomous robots. *Robotics and Autonomous Systems*, 55 (9), 750-759, 2007.

[9] C. Rothkopf, D.H. Ballard, and M. M. Hayhoe. Task and context determine where you look. *Journal of Vision*, 7:1–20, 2007.

[10] M. Schlesinger, D. Amso, and S.P. Johnson. The neural basis for visual selective attention in young infants: A computational account. *Adaptive Behavior*, 15(2):135–148, 2007.

[11] J. Vitay, N.P. Rougier, and F. Alexandre. A distributed model for spatial visual attention. *In: Wermter et al. (Eds.): Biomementric Neural Learning, LNAI 3575, Springer*, pages 54–72, 2005.