

# A head–neck–eye system that learns fault-tolerant saccades to 3-D targets using a self-organizing neural model

Narayan Srinivasa<sup>a,\*</sup>, Stephen Grossberg<sup>b</sup>

<sup>a</sup> Department of Information and System Sciences, HRL Laboratories LLC 3011, Malibu Canyon Road, Malibu, CA – 90265, United States

<sup>b</sup> Department of Cognitive and Neural Systems, Center for Adaptive Systems and Center for Excellence for Learning in Education, Science and Technology, Boston University, 677 Beacon Street, Boston, MA – 02215, United States

## ARTICLE INFO

### Article history:

Received 1 February 2007

Revised and accepted 31 July 2008

## ABSTRACT

This paper describes a head–neck–eye camera system that is capable of learning to saccade to 3-D targets in a self-organized fashion. The self-organized learning process is based on action perception cycles where the camera system performs micro saccades about a given head–neck–eye camera position and learns to map these micro saccades to changes in position of a 3-D target currently in view of the stereo camera. This motor babbling phase provides self-generated movement commands that activate correlated visual, spatial and motor information that are used to learn an internal coordinate transformation between vision and motor systems. The learned transform is used by resulting head–neck–eye camera system to accurately saccade to 3-D targets using many different combinations of head, neck, and eye positions. The interesting aspect of the learned transform is that it is robust to a wide variety of disturbances including reduced degrees of freedom of movement for the head, neck, one eye, or any combination of two of the three, movement of head and neck as a function of eye movements, changes in the stereo camera separation distance and changes in focal lengths of the cameras. These disturbances were not encountered during motor babbling phase. This feature points to general nature of the learned transform in its ability to control autonomous systems with redundant degrees of freedom in a very robust and fault-tolerant fashion.

© 2008 Elsevier Ltd. All rights reserved.

## 1. Introduction

In humans, there are basically two types of eye movements. The most common eye movement is to keep the gaze affixed. These gaze holding movements are the result of compensating for head movements by moving the eyes in an equal and opposite amount to the direction of the head movements. These movements are either driven by the balance organs of the inner ear (called the vestibule-ocular reflexes or VOR), or alternatively they can be driven by the retinal image motion in a feedback loop (called the optokinetic responses or OKR). The other main class of eye movement comes about because the fovea (center high resolution portion of the retina), has a high concentration of color sensitive photoreceptor cells called cone cells. The rest of the retina is mainly made up of monochrome photoreceptor cell called rod cells, which are especially good for motion detection. By moving the eye so that small parts of a scene can be sensed with greater resolution, body resources can be used more efficiently. The eye movements disrupt vision and hence we have evolved to make these movements as

fast, and therefore as short in duration, as they can possibly be; they are called *saccades*. In addition, since we have two eyes, they need to be coordinated so that images of an object fall on exactly the same parts of the two retinae.

A feature of the human eye system is that the total degrees of freedom available for use to perform coordinated eye movements is far greater than that required to fixate or saccade to 3-D targets. These redundant degrees of freedom are exploited by the human brain to derive flexible ways to saccade to 3-D targets. There have been several attempts to develop robotic camera systems that can saccade to 3-D targets (Aloimonos, 1990; Batista, Peixoto, & Ara'ujo, 1997; Batista, Dias, Araujo, & Almeida, 1995; Brown & Coombs, 1993; Dias et al., 1997; Murray, Bradshaw, MacLauchlan, Reid, & Sharkey, 1995; Sharma, 1994; Srinivasa & Ahuja, 1998; Srinivasa & Sharma, 1997, 1998; Wei & Ma, 1994). The novel aspect of this work is that we demonstrate a fully self-organized approach to learning how to perform saccadic control despite redundancies in the system. Furthermore, we also demonstrate that such a control system offers robustness to various disturbances that the system has not experienced *a priori*. This feature of our work has seldom been demonstrated in previous work for saccade control. In this paper, our goal is to develop a self-organized learning process that can enable a robotic head–neck–eye camera system

\* Corresponding author. Fax: +1 310 317 5958.

E-mail address: [nsrinivasa@hrl.com](mailto:nsrinivasa@hrl.com) (N. Srinivasa).

with 12 degrees of freedom to saccade to 3-D targets. Here the number of degrees of freedom available is more than the space in which the goal is specified (4-D coordinates – the desired location of the image coordinates of a 3-D target). This system is an example of a *motor equivalent* system (Bullock, Grossberg, & Guenther, 1993) because there are several possible alternatives to saccade to a given 3-D target. These alternatives are derived from various combinations of the redundant degrees of freedom of the head–neck–eye camera system. A redundant head–neck–eye camera system generates self-consistent signals between vision and motor systems via action perception cycles. These signals are then used by a self-organizing neural model to learn how to control the head–neck–eye movements to saccade to 3-D targets in a robust and fault-tolerant fashion.

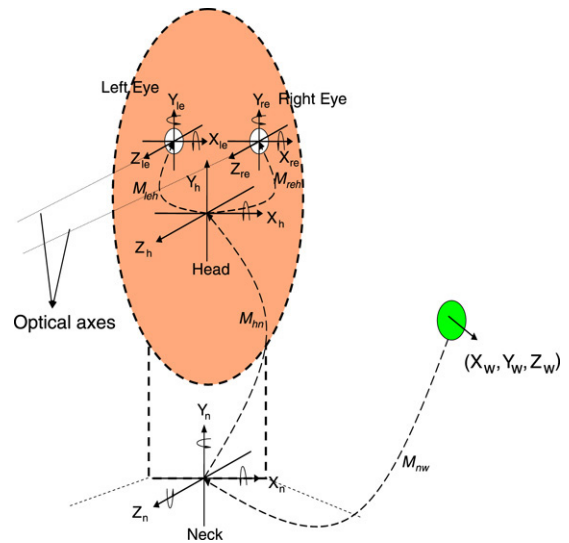
The paper is organized as follows. The next section will introduce the notion of action perception cycles. This will be followed in Section 3 with an introduction to the head–neck–eye camera model. In Section 4, the self-organized learning process using the head–neck–eye camera model will be outlined. The next section will highlight the performance of the learned transform to saccade to new 3-D targets. In this section, the results on testing the saccade system for robustness to various disturbances and constraints will also be provided. In Section 6, the basis for model robustness will be provided. The biological plausibility of the model will also be discussed here. In Section 7, conclusions will be provided followed by details of the head–neck–eye model in the Appendix.

## 2. Action-perception cycles

Perceiving without acting is not common. For example, scrutinizing an object visually presupposes saccades at it and sometimes involves moving the head or even the whole body. Similarly, to accurately localize a sound source it becomes necessary to move head and ears towards the sound source. Acting without perceiving seldom makes sense; after all, actions defined as goal-directed behavior, aim at producing some perceivable event – the goal. Performing an appropriate action requires perceptual information about suitable starting and context conditions and, in the case of complex actions, about the current progress in the action sequence. Thus, perception and action are interlinked or interdependent.

There are several behavioral repertoires in which this interdependency is manifested in humans and other species. In the simplest form, a behavior is triggered by the present situation and reflects the animal's immediate environmental conditions. This type of behavior is often referred to as *stimulus-response reflexes*. A good example of this type of behavior in humans is provided by the orientation reflex, which we exhibit when encountering a novel and unexpected event. On the one hand, this reflex inhibits ongoing actions and tends to freeze the body – a stimulus triggered response. At the same time, it also draws attention towards the stimulus source by increasing arousal and facilitating stimulus-directed body movements. This interdependency between stimulus and response creates an *action perception cycle* (Piaget, 1963) wherein a novel stimulus triggers actions that lead to a better perception of itself or its immediate environmental condition and the cycle continues.

Human behavior is much more flexible than exclusive control by stimulus-response cycles. One of the hallmarks of human capabilities is the ability to *learn* new relations between environmental conditions and appropriate behavior during action perception cycles. This learning process provides an enormous gain in flexibility for an individual by creating the ability to adapt to environmental changes. Not only we learn to react to particular environmental conditions and situations in a certain way, we



**Fig. 1.** A schematic of the head–neck–eye camera model is shown here. The details of the notations and transformations between the various coordinate frames are provided in the Appendix.

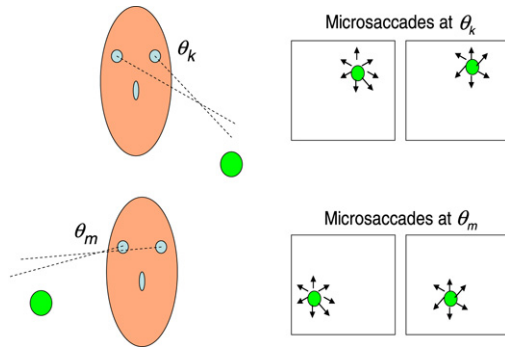
also can unlearn what we have acquired and learn new relationships between situations and actions. In fact, it is known (Barto & Sutton, 1982; Barto, 1995; Grossberg, 1972, 1982) that humans can condition their behaviors based on rewards/punishments that they may receive while interacting with its environment (also known as reinforcement learning). Furthermore, the ability to implement and switch between learned behaviors forms the basis of highest degrees of behavioral flexibility. It is the goal of this paper to study how action perception cycles could play a part in enabling a head–neck–eye camera system to *learn to saccade* to 3-D targets in a manner that is robust and tolerant to new disturbances and unexpected situations.

## 3. The head–neck–eye camera model

The human visual system is an active vision system that can be controlled by the brain in a deliberate fashion to extract useful information about the environment. The head–neck–eye camera model used in this work is an abstracted version of the human active vision system. It consists of a pair of cameras (eyes) mounted on a head and the whole system is supported by neck (refer to Fig. 1). The head–neck–eye system consists of 12 degrees of freedom: each eye has 8 degrees of freedom – a single tilt and two independent pan for rotation, two degrees of freedom for controlling the focal length of each camera, two degrees of freedom for the retinal image center for each camera and an additional degree of freedom in controlling the baseline distance between the two cameras; the neck has 3 degrees of freedom – tilt, pan and yaw (shoulder-to-shoulder) for rotation; and finally the head has one degree of freedom – tilt for rotation. During action perception cycles, only the rotational degrees of freedom (or extrinsic parameters) were exercised. The imaging parameters (or intrinsic parameters) were fixed. The kinematics that describes the imaging of a 3-D object in both the eyes as a function of the extrinsic parameters and intrinsic parameters is provided in the Appendix.

## 4. Self-organized learning of saccades via action perception cycles

During action perception cycles for learning to saccade, the head–neck–eye camera system is setup to look at 3-D targets in its visible space for various joint configurations of camera system.



**Fig. 2.** A couple of steps during the action perception cycles are shown here for illustration. The camera joint configuration denoted by  $\theta_k = \{\alpha_N, \beta_N, \gamma_N, \alpha_H, \alpha_e, \beta_{Le}, \beta_{Re}\}$  corresponds to the seven d.o.f. of the camera. For each such camera configuration, the camera performs a set of actions – microsaccades and the resulting perception is a set of translations of the 3-D target (green sphere shown here) in various directions. Similar actions are performed at the second joint configuration  $\theta_m$ . This process is repeated during the action perception cycles where the 3-D target is placed uniformly spanning the visible space of the head–neck–eye camera system.

Each head–neck–eye camera joint configuration  $\theta$  corresponds to a unique joint position of each of the seven degrees of freedom (three neck –  $\alpha_N, \beta_N, \gamma_N$ , one head –  $\alpha_H$  and three eye –  $\alpha_e, \beta_{Le}, \beta_{Re}$  rotation angles). It should be noted that we will also use the notation  $\theta_i$  (for  $i = 1, \dots, 7$ ) to refer to the seven rotational degrees of freedom interchangeably throughout the rest of the paper. These joint configurations represent the context for the learning system. At each joint configuration, the camera performed a set of microsaccades wherein the joints of the camera system was exercised to move in small increments. These actions resulted in translation of image of the 3-D target in various directions within the image plane of both the cameras (refer to Fig. 2).

The differential relationship between the spatial directions of target image in the retina to the joint rotations of the head–neck–eye camera system as a result of the microsaccades for a given context  $\theta$  during action perception cycles is a linear mapping. Our system learns this mapping in a self-organized fashion (as described below). For a redundant system like the head–neck–eye camera system used in this paper, this linear mapping is a one-to-many function. This implies that there exists several possible linear combinations of solutions from spatial directions of the target image in the stereo camera to head–neck–eye camera joint angle changes that can generate a single image space trajectory of the target that is continuous in joint space and correctly directed in the 4-D space (two direction vectors corresponding to the stereo pair directed towards the retinal center for each eye – i.e., to saccade). For example, to look at a 3-D target, it is possible to just move only the eyes with respect to  $\theta$  in order to fixate on the target provided the target is visible and the eye joints are within the physical limits of its joint rotation space or joint space. At the same time, it may also be possible to use some of the other joints including the head and neck in addition to the eye joints to fixate on the same 3-D target. Joint space continuity is ensured because all solutions are in the form of joint angle increments with respect to the present fixed joint configuration  $\theta$  of the head–neck–eye camera system.

This synchronous collection of increments to one or more joint angles of the head–neck–eye camera system is called a *joint synergy* (Bullock et al., 1993). During the self-organized learning process, the head–neck–eye motor system learns to associate a finite number of joint synergies to the spatial direction of image movements in the stereo camera when these synergies are activated for a given  $\theta$ . During performance, a given desired movement direction of the target (in the case of saccades the

desired direction is toward the retinal image center) can be achieved by activating in parallel any linear combination of the synergies that produces that image movement direction. This simple control strategy leads to motor equivalence when different linear combinations are used on different movement trials. The self-organized learning process utilizes a self-organizing neural model that will now be described.

The neural architecture for learning to saccade to 3-D targets is shown in Fig. 3. The network consists of four types of cells. The *S* cells encode the spatial directions of the target when the camera is either babbling or performing a learned saccadic movement. The *V* cells encode the difference between weighted inputs from the direction cells *S* and the *R* cells that encode the joint rotation directions or increments of the head–neck–eye system. The network adapts the weights *Z* between the *S* cells and the *V* cells based on the difference of activity between the spatial directions of the target motion in the cameras and joint rotations of the head–neck–eye camera system. We have adopted the VAM learning approach for this learning process (Gaudio & Grossberg, 1991; Srinivasa & Sharma, 1998). During learning, the *V* cell activity drives the adjustment of weights. This process is akin to learning the pseudo-inverse of the Jacobian between spatial directions to joint rotations. During performance, the learned weights are used to drive the *R* cells to the desired increment in joint rotation.

In order to ensure that the correct linear mapping is learned and the motor equivalence can be addressed, the learning process has to account for the joint configuration  $\theta$  under which learning of the mapping takes place. We refer to each joint configuration as the context and a set of neurons or *C* cells that encode various contexts that span the joint space of the head–neck–eye camera system as the context field (Bullock et al., 1993; Fiala, 1994). The *C* cell neuron in the context field strongly inhibits the *V* cells allocated for that context (refer to Fig. 3). When a *C* cell neuron in the context field is excited due to the head–neck–eye system being in the appropriate configuration, it momentarily inhibits the *V* cells allocated for that context and this allows the learning process to adapt the weights in a manner that enables the computation of the correct linear mapping. An overall flowchart in Fig. 4 summarizes the neural network model and the various steps during both the training and performance phases. Table 1 summarizes the network equations of the model for these two phases.

## 5. Computer simulations

The head–neck–eye system used in this work is a seven-dof for angular position control (extrinsic parameters) and five degrees of freedom (change in focus and separation in baseline between the eyes and location(s) of the retinal center in the stereo images). The system has more degrees of freedom than required to saccade to a 3-D target thereby making the system redundant.

### 5.1. Learning

The system was trained using a single 3-D target (a sphere) during action–perception cycles to learn appropriate weights to saccade to the target (refer to Fig. 2). The process begins by motor-babbling where the head–neck–eye camera system performs random eye movements that exercise all the seven rotational degrees of freedom. These movements are two types. First, it performs a gross movement wherein the camera moves to distinctly different camera configurations. At each of these camera configurations, a set of microsaccades were performed and the direction-to-rotation transform was (as described earlier) learned at each camera configuration or context.

The weights *Z* were initialized to zero and subsequently adapted (refer to Table 1) on a context basis as and when the

**Fig. 3.** The neural architecture for learning saccades to 3-D targets is shown here. The network shows the *S*, *V* and *R* cells and how their interactions during perception, learning and action cycles enable the network to adapt the weights *z* to learn to saccade. During performance, the network can integrate the *R* cell activity to produce new joint configurations that move the head–neck–eye camera system to saccade to a 3-D target.

**Table 1**  
The network equations during learning and performance phases are listed for various cells and computations

Network cells/computations	Learning phase	Performance phase
Direction cells ( <i>S</i> )	$\frac{ds_j}{dt} = -\lambda s_j + (1 - s_j) s_j - \sum_{l \neq j} s_l$	$\frac{ds_j}{dt} = -\lambda s_j + (1 - s_j) d_j - s_j \sum_{l \neq j} d_l$
Difference cells ( <i>V</i> )	$\frac{dv_{jk}}{dt} = -\alpha v_{jk} + \sum_j z_{ijk} s_j - R_i$	$\frac{dv_{jk}}{dt} = -\alpha v_{jk} + \sum_j z_{ijk} s_j - R_i$
Joint rotation cells ( <i>R</i> )	$\frac{dR_i}{dt} = \delta (r_i - R_i)$	$\frac{dR_i}{dt} = \delta (-R_i + \sum_k v_{ik})$
Weights ( <i>Z</i> )	$\frac{dz_{ijk}}{dt} = \gamma v_{ik} s_j$	No learning
Joint configuration ( $\theta$ )	$\theta_i = \theta_i^{\min} + \text{rand}^*(\theta_i^{\max} - \theta_i^{\min}) \quad i = 1, \dots, 7$ $\theta_1 = \alpha_N; \theta_2 = \beta_N; \theta_3 = \gamma_N; \theta_4 = \alpha_H$ $\theta_5 = \alpha_E; \theta_6 = \beta_{Le}; \theta_7 = \beta_{Re};$	$\frac{d\theta_i}{dt} = -\eta \theta_i + G[R_i] + \theta_{old}; \quad i = 1, \dots, 7$
Context cell selection	$C_i = \ \theta_c - \theta_i\ $ $k = \max_k(C_k)$	$C_i = \ \theta_c - \theta_i\ $ $k = \max_k(C_k)$

The direction cells *S* for the learning phase obey a center-surround type of computation (Grossberg, 1988) where the direction inputs *s* are normalized by the *S* field. For the performance phase, the *S* cells are normalized in a similar fashion. However, the desired direction to image center *d* is used in the computation. The *V* cell computations are based on the difference between the weighted direction inputs and joint rotation increments. The *R* cell computations during learning phase is based on the motor babbling increment *r* whereas during performance is derived from the population of *V* cell activity whose contexts *k* are relevant to the head–neck–eye configuration. The learning updates of *z* are based on the activity of the *V* and *S* cells. During learning phase the joint configurations are computed randomly as part of the motor babbling phase while they are updated during performance phase by the product of the GO signal which controls the speed of the camera movement. The context cell selection during learning and performance is based on selecting the cell *k* that best matches (*L*<sub>2</sub> norm or Euclidean distance) the current head–neck–eye camera configuration.

appropriate context node became active. In our simulations, the range of the joints used for the seven rotational degrees of freedom and the number of discretized zones for each angle is listed in Table 2. This discretization process yielded a total of 77175 contexts cells (5 × 5 × 3 × 3 × 7 × 7 × 7). At each camera configuration a total of 100 randomly generated microsaccades were performed to compute the direction-to-rotation transform for that context. The sequence of steps during the learning phase is summarized in Fig. 4(i). The various parameters used during the learning phase for the equations listed in Table 1 are: λ = 0.01, α = 10.0, δ = 32.0 and γ = 8.0. All simulations were performed using the 4th order Runge–Kutta ODE solver with a time step of 0.001. The total duration of learning was 2.5 h on Dell XPS computer with a 2 GB RAM.

### 5.2. Performance

The performance phase begins when the learning phase is completed. It should be noted that this does not have to be the case in general. The nature of the learning algorithm (i.e., VAM learning – Gaudio and Grossberg (1991)) allows training and performance phases to be interleaved. In order to saccade to a visible target, the desired spatial direction from the current location of target in the stereo images to the image centers is computed. The context or camera configuration of head–neck–eye system is used to access the appropriate direction-to-rotation transform. This transform is used to compute the desired joint angle increments of the head–neck–eye camera system. These increments are finally integrated over time to saccade to the 3-D















