

Adaptive image stabilization: a need for vision-based active robotic agents

Francesco Panerai

LPPA, Collège de France
11 pl. M. Berthelot
75005 Paris
panerai@college-de-france.fr

Giorgio Metta, Giulio Sandini

LIRA Lab, DIST - University of Genova
Viale F. Causa 13
16145 Genova, Italy
{pasa, giulio}@lira.dist.unige.it

Abstract

This paper describes a framework which enables to develop an adaptive image stabilization mechanism for robotic agents (RAs) with moving eyes. In analogy with solutions found in natural systems, our RA exploits two sensory systems (inertial and visual) to capture movement of the head in three-dimensional space. While the RA interacts with the environment, a neural network (NN) learns to transform the movement related sensory signals into compensatory motor commands to stabilize gaze. The system achieves satisfactory stabilization performance through an unsupervised, visually driven learning scheme. Two experiments show stabilization performance measured directly on the image plane. The adaptive properties of the stabilization mechanism are discussed in relation to the external world (i.e. the environment) and to the available computational resources of the artificial system.

1. Introduction

In animal species with moving eyes, image stabilization is obtained efficiently through reflex compensatory movements (Wilson and Melvill Jones, 1988). Intriguingly enough, in many cases the reflex eye movements are already present in the early period of the life. Also in neonates, the VOR stabilization reflex is clearly present (Finocchio DV et al. 1991). In this early period of their life, infants rely predominantly on “pure vestibular” reflex (e.g. VOR gain is on the average 1.0) which reduces reliance on a poorly developed optokinetic and smooth pursuit system. The presence of an early stabilization reflex is thought to contribute to the infant’s early visual processing development. On the contrary, in the adulthood, the maturation of the latter two systems helps the VOR provide perfect ocular stabilization and leads to a reduction of the average VOR gain (e.g. gain reduces to about 0.59). Although the development of VOR and its interaction with

the developing optokinetic and smooth pursuit eye movements are not completely understood (Shupert, 1988) it is clear that important changes take place over time, adapting the dynamic characteristics of this stabilization mechanism (Weissman, 1989), (Ornitz, 1985). The brain system responsible for this stabilization mechanism is the vestibulo-ocular reflex circuitry (VOR). One remarkable aspect of such circuitry is its plasticity: it adapts/reacts to any change which degrades image stabilization performance. Whatever the change, an error signal is created which informs the brain that the VOR is not working properly. As a result the system recalibrates itself. The recalibration can occur over the course of hours to days and at the end of it, the newly calibrated system has stored a new set of “stabilization parameters”. It has been demonstrated that VOR recalibrates when it is inaccurate and images move across the retina during head turns (Miles and Fuller, 1974). This type of motor learning has been studied from different perspectives and model of the learning process and sites of learning have been proposed (Lisberger, 1988). The neural region that seems responsible for these kind of recalibration is the cerebellum (Lisberger, 1998). The gain of the VOR reflex is nominally 1.0 and it is kept close to this value by a parametric-adaptive control system (Shelhamer et al. 1992).

In robotics, image stabilization techniques exploiting compensatory camera movements have received little attention so far (Panerai and Sandini, 1998), (Panerai et al. 2000), (Shibata and Schaal, 1999). For a decade a growing number of studies have concentrated on active control of camera movements (Aloimonos et al. 1988), (Krotkov, 1989), (Ballard and Brown, 1992), (Sharkey et al. 1993), (Capurro et al. 1997), (Nordlund and Uhlin, 1995), (Rougeaux, 1999) and sophisticated binocular machines have been successfully implemented (Uhlin et al. 1995), (Murray et al. 1995). On the other hand, although robust performance has certainly been obtained when the RAs play as static observers in “structured” environment, the performance in many cases is completely disrupted when operating in “disturbed”, non-structured conditions. When

interaction with the environment becomes an important issue for the RA, it is worth addressing the question on how visual performance might be kept unchanged in presence of external sources of disturbances. Either these being simply produced by navigating through a rough terrain, or otherwise due to self generated movements, what could possibly be the solution to avoid the degradation of robot's visual functionalities? In previous work we have shown that an oculomotor control architecture integrating inertial and visual sensory information avoids degradation of visual performance when external motion/disturbances occur (Panerai and Sandini, 1998), (Panerai et al. 2000).

In this work we propose a solution which enables a RA to develop autonomously adaptive image stabilization functionalities. A Growing Neural Gas (GNG) network receives as inputs two motion related cues (i.e. the vestibular information and the retinal motion information) and adjusts its parameters to generate optimal stabilization motor commands. The result of the adaptive learning scheme is the construction of a sensory-motor map which codes the compensatory stabilization reflexes. The only basic assumptions made at the beginning of the learning process are that the RA should be able to compute some form of retinal slip and have access to the inertial information.

In section 2 the variables chosen to control ocular movements are described. Section 3 treats of the network model and the learning paradigm. Section 4 focuses on the advantages of visuo-inertial stabilization for developing robotic agents and finally, section 5 shows the performance results in image stabilization obtained with this approach.

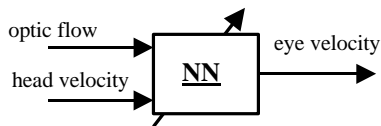


Figure 1: **The experimental setup.** Top: the robot system is a 5 DOF binocular head integrating an artificial vestibular system (i.e. the white bloc in the centre of the picture). Bottom: a block diagram of the NN, its inputs (optic flow, head velocity) and output (eye velocity).

2. Oculomotor control for image stabilization

The movement of a camera in three-dimensional space produces an optic flow (Sundareswaran, 1991) that under simplified hypotheses (i.e. center of the image plane, motion limited to one rotational and one translational components) can be written as:

$$u_0 = f_x \left[-\frac{T_x}{Z(0,0)} - W_y \right] \quad (1)$$

where T_x represent the translation along the fronto parallel image plane, W_y the rotation around a vertical axis, f_x the focal length and $Z(0,0)$ the distance to the fixation point. By substituting the term W_y with the sum of the head velocity Ω_y and the eye velocity with respect to head w_y we have:

$$\frac{u_0}{f_x} = -\frac{T_x}{Z(0,0)} - (\Omega_y + w_y) \quad (2)$$

Eq. (2) tells us that the Ω_y and w_y variables, if directly measurable by the RA, are the most indicated to learn to generate compensatory eye movement w_y for image stabilization purposes. We have then envisaged the use of these two variables as independent inputs of a neural network which learns to approximate a control surface representative of the optimal compensatory command w_y (optimal in the sense it minimizes the residual optic flow in the image plane).

3. The neural network model

The neural network architecture is built on a GNG-Soft model (Metta, 1999). It combines two network models, namely, the Growing Neural Gas (GNG) model and SoftMax function network. The resulting hybrid architecture has several advantages. First, the effectiveness, typical of the GNG, in distributing the units within the multi-dimensional input space. Second, the "optimal" approximation and interpolation properties of SoftMax functions networks. Third, an interesting (with relation to our task) self-tuning capability. Structurally, the network consists of a single layer of processing elements (PEs), each characterized by a receptive field-like structure. The single PE's response can be described analytically by the following expression:

$$U_i(\mathbf{x}, c_i) = \frac{G(\|\mathbf{x} - c_i\|)}{\sum_j G(\|\mathbf{x} - c_j\|)} \quad (3)$$

where $G(\cdot)$ is a Gaussian function, $\mathbf{x} \in \mathfrak{R}^N$ is the input to the network and c_i the receptive field positions. The output of the network is the linear combination of a number of PEs.

Analytically we have:

$$g(\mathbf{x}) = \sum_i v_i U_i(\mathbf{x}, c_i) \quad (4)$$

where i extends to the number of units mapping the input space and the parameters v_i are the weights of the output layer ($g(\mathbf{x}) \in \mathfrak{R}$). The network parameters which will be tuned during the unsupervised learning process are: c_i (the function's centers), v_i (the weights of the output layer), \mathbf{s}_i (the standard deviation of each Gaussian functions). The learning process consists of incrementally adjusting these parameters to improve/reduce a predefined "performance index" over time.

3.1 Learning principle and learning scheme

In our framework the input space of the network is two-dimensional. It is defined by the instantaneous angular velocity of the head (i.e. Ω_y) and by the instantaneous optic flow (i.e. u_0) measured on each camera image plane. To adjust the network parameters, we have chosen a performance index measuring the instantaneous component of the residual optic flow (ROF) at the center of the image plane (u_0). The tuning of the network parameters has the goal of minimizing the ROF on the image plane, that is:

$$\min_{n_i, c_i} \left[-\frac{T_x}{Z(0,0)} - (\Omega_y + \sum_i v_i U_i(\mathbf{x}, c_i)) \right] \quad (5)$$

Self-normalizing Hebbian rules are used to modify the PEs centers c_i , the weights v_i of the output layer according to the following:

$$\Delta c_i = \mathbf{h}_1(\mathbf{I} - c_i) U_i \quad \Delta v_i = \mathbf{h}_2(\mathbf{V} - v_i) U_i$$

A heuristic criterion is used in order to tune the Gaussian's variances \mathbf{s}_i . A description of this can be found elsewhere (Metta, 1999). On the other hand, in order to carry out the minimization, the weights of the output layers are modified as follows:

$$\Delta n_i = \mathbf{h}_2 \left(\sum_j n_j U_j(\mathbf{x}, c_j) + u_0 - n_i \right) \cdot U_i(\mathbf{x}, c_i) \quad (6)$$

that is the target output is shifted by the quantity u_0 from the current network output. Whenever, stabilization is perfect (i.e. $u_0=0$), no adjustment is necessary and in fact $\Delta n_i \approx 0$. It is worth stressing that time, which is not

explicitly indicated in equation (6), plays a fundamental role in this schema. In fact, the optic flow used as input to the network is actually one time step before of that used as stabilization measure. That is, the measure of the network performance can be obtained only one step after the network has been used to generate a motion command. A delay line in Figure 2 indicates this last point.

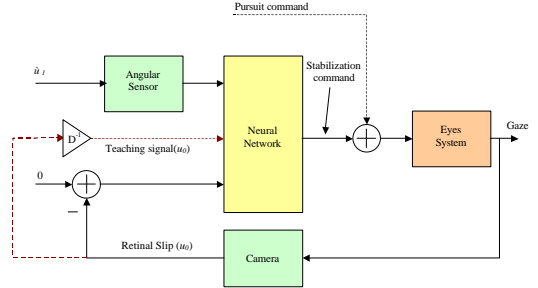


Figure 2 **Motor learning scheme.** The inertial information (angular sensor) and the retinal slip (optic flow) are combined by the NN. The teaching signal is the optic flow itself, which has to be minimized for stabilization to be effective.

4. The "visuo-inertial" approach and the available computational resources

In biological systems, from a developmental point of view, one of the advantages obtained by integrating visual and inertial information, is that of adapting to the available computational resources at a given time. In the introduction we have mentioned the example of neonates, which possess in their early age a pre-developed pure vestibular reflex (Finocchio DV et al. 1991). In fact, during this time the brain circuitry devoted to processing visual information is not yet efficient. The OKN and the smooth-pursuit systems in particular are poorly functional. On the other hand, because of the developmental issues, coordination between eye movements and visual processing is of paramount importance in this period (think for example to the development of binocular visual processing) (Held et al. 1996). For a developing system, it is therefore reasonable to approach a partial and temporary solution consisting of a hard-wired stabilization reflex which helps to stabilize the visual world and at the same time facilitates retinal correspondence between the two eyes. Inspired by this idea, we performed two experiments in which the NN learns to generate the stabilization reflex exploiting different amounts of inertial and visual information. The same exact form of interaction with the environment is engaged in the two cases (a sequence of fixed amplitudes and random amplitudes movements). In the first case, the parameters of the NN are initialized to mimic a sort of immature (i.e. pure vestibular) reflex. In such condition, the learning scheme described in

section 3.1 leads the NN to develop the sensory motor map depicted in Figure 3. In the second experiment, the NN's parameters are initialized to assign the RA a more "mature" (i.e. visuo-inertial) stabilization reflex. This assumption leads to the development of the sensory motor map depicted in Figure 4.

The comparison of the two maps shows appreciable differences in terms of the input domain and the shape of the control surfaces. In Figure 3, for example, the domain of the ROF is limited to the $(-0.1, 0.1)$ interval. On the contrary, the domain extends to a larger interval $(-0.6, 0.6)$ in Figure 4. Note also that the distribution of the NN's units is different in the two cases: it is condensed in the centre in the "pure vestibular" case, while it spreads toward the boundaries of the ROF domain in the "visuo-inertial" case. Indeed, from the RA's point of view, the first case (i.e. the reflex being predominantly vestibular) represents a less demanding requirement in terms of visual motion processing. On the other hand, the second case, requires the RA to dedicate more processing to image motion analysis since performance in this case strongly depends on the ROF measurements.

The two experiments highlight the interesting property of the NN "visuo-inertial" stabilization approach which well adapts to the amount of computational resources available in the RA at a given time. Note that in a RA computational resources are always limited. More importantly, they should be appropriately distributed in order to be sufficient to

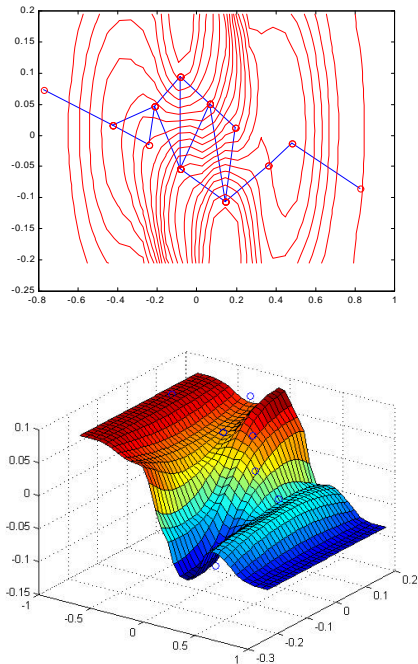


Figure 3 – Sensory motor map of the "pure vestibular" reflex. Top: the input domain of the neural network (flow, inertial) and the network units (small circles) distribution. Bottom: the control surface interpolated by the network units.

process the different type of sensory information (visual and inertial) for real-time interaction with the environment. Therefore a straight hard-wired initial solution, although sub optimal, might be advantageous for the RA to solve the problem of image stabilization in the context of a developmental framework. The "smoothness" of the control surface is also worth it considering. In the two cases it is different: In the second experiment in particular, the degree of smoothness of the sensory motor map is much more evident than in the first case (although both surfaces are optimal in terms of the ROF criterion specified in the learning scheme). Therefore, if one of the RA's concern is to minimize "energy consumption" during compensatory ocular movements, the former solution is, once again, only sub optimal.

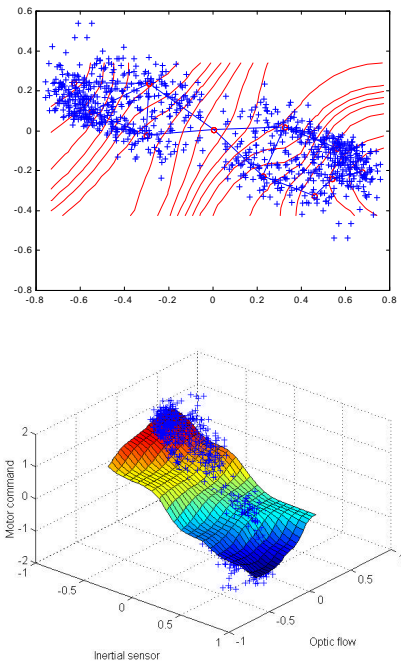


Figure 4 – Sensory motor map of the "visuo-vestibular" reflex. Top: the input domain of the neural network (flow, inertial) and the network units (small circles) distribution. Bottom: the control surface interpolated by the network units.

In our view, time-varying learning schemes might represent an interesting approach to be investigated within the context of developmental robotics, always having in mind that priority for a RA should be devoted to getting "up and running" in the shortest time.

5. Stabilization performance

The stabilization mechanism enables the RA to generate correct compensatory eye movements from the very beginning. In Figure 5 the compensatory behavior is described during an initial and advanced phase of learning. In the two plots, the following measurements are represented: the external rotational stimulation as measured by the inertial sensor, and the motor command synthesized

by the learning network. Note how the motor command increases consistently its amplitude during the first 900 cycles of learning. During the advanced phase of learning (shown in Figure 5, bottom) the motor response still grows, but at a slower rate. Within our framework, learning has been performed using different stimuli, changing the amplitude and the frequency of the externally imposed movement. Random stimulation have also been used to simulate a persistent external disturbance. In all cases convergence toward a stable behavior have been achieved.

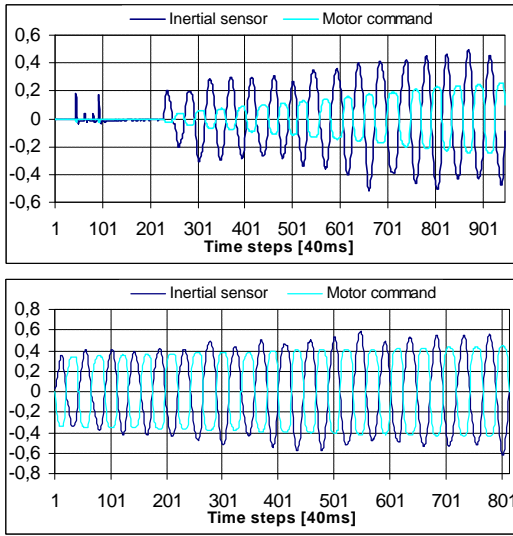


Figure 5 – **Initial and advanced phase of the motor learning** An external source produces a repetitive rotational movement whose amplitude increases in time. The NN the two motion signals and generates a compensatory command to minimize retinal slip. The trajectories shows the inertial sensor and the compensatory command (normalized). Time is expressed as control cycles (40ms). Top: during the initial phase the motor command steadily increases showing that the compensatory reflex is not effective yet. Bottom: in the advanced phase compensatory motor response is still growing, even if at a slower rate.

Direct measurements on the image plane is of paramount importance if one wishes to evaluate the impact of external disturbances on the visual functionalities of the system. Therefore, stabilization performance has been evaluated by means of optic flow techniques through first order estimates (see (Capurro et al. 1997) for the algorithm). In all performance measurements, the same sensory motor map is used to generate the compensatory camera movements. Figure 6 shows on a normalized scale the inertial measurement (i.e. the angular velocity) and the image slip (u_0 component of the optic flow) corresponding to two different external motions (stimuli characteristics are respectively 0.3 Hz, 18 deg/s amplitude and 0.6 Hz, 81 deg/s amplitude). We have evaluated numerically the amount of ROF. In correspondence to the maximum peak velocity of each stimulus, the ROF is less than 1

pixel/frame. It is worth noting that, with the second stimulus, the frequency and the amplitude of the external movement change substantially (i.e. frequency roughly doubles and amplitudes increases four times approximately), but the amount of retinal slip is still very limited.

6. Conclusion

We have described a framework for the development of oculomotor stabilization reflexes in vision-based active robotic agents (RAs). Sensory information about the RA’s self-motion are obtained using an artificial vestibular apparatus and a basic motion detection algorithm. The motion cues are integrated by means of an efficient neural controller and used to generate compensatory camera movements. An unsupervised learning scheme enables the RA to build a sensory motor map which transforms self motion signals into compensatory motor commands. The learning scheme is efficient in adapting the network parameters and becomes effective after a short training period. Interesting enough, different initialisation of the neural controller enable to describe emerging property of the “visuo-inertial” stabilization approach: adaptation to the available computational resources. This is true for artificial as well as natural systems. Two experiments prove that stabilization is indeed achieved using this approach.

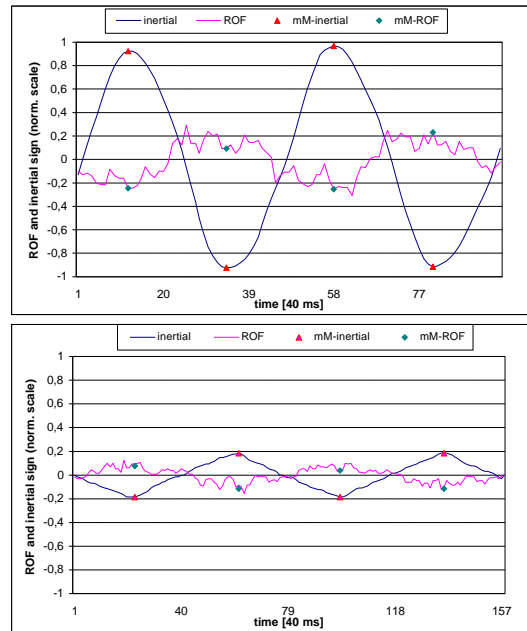


Figure 6 – Residual optic flow (ROF) during stabilization. Top: a normalized scale of the ROF measured (norm. fact. 5 pixels/frame) and the angular velocity component of the external movement (max 90 deg/s). For a rotational movement of about 80 deg/s, ROF is bounded to less than 1 pixel/frame. Bottom: ROF is bounded here to 0.5 pixel/frame (0.1*5 pixels/frame) for a rotational movement of 18 deg/s.

References

- Aloimonos, J., Weiss, I. and Bandopadhyay, A. (1988) Active Vision. *International Journal on Computer Vision* **1**, 335-356.
- Ballard, D.H. and Brown, C.M. (1992) Principles of Animate Vision. *Computer Vision Graphics and Image Processing* **56**, 3-21.
- Capurro, C., Panerai, F. and Sandini, G. (1997) Dynamic Vergence using Log-polar Images. *International Journal on Computer Vision* **24**,
- Finocchio DV, Preston KL and Fuchs AF (1991) Infant eye movements: quantification of the vestibulo-ocular reflex and visual-vestibular interactions. *Vision Res* **31**, 1717-1730.
- Held, R., Thorn, F., Gwiazda, J. and Bauer, J. (1996) Development of binocularity and its sexual differentiation. In: Vital-Durand, F., Atkinson, J. and Braddick, O.J., (Eds.) *Infant Vision*, pp. 265-274. Oxford University Press]
- Krotkov, E.P. (1989) *Active computer vision by cooperative focus and stereo*, Springer-Verlag edn.
- Lisberger, S.G. (1988) The neural basis for motor learning in the vestibulo-ocular reflex in monkeys. *Trends Neurosci* **11**, 147-152.
- Lisberger, S.G. (1998) Physiological basis for motor learning in the vestibulo-ocular reflex. *Otolaryngol Head Neck Surg* **119**, 43-48.
- Metta, G. (1999) Babyrobot: a Study an Sensory-motor Development. LIRA Lab, DIST, University of Genova. Ph.D. Thesis.
- Miles, F.A. and Fuller, J.H. (1974) Adaptive plasticity in the vestibulo-ocular responses of the rhesus monkey. *Brain Res* **80**, 512-516.
- Murray, D.W., Bradshaw, K.J., McLauchlan, P.F., Reid, I.D. and Sharkey, P.M. (1995) Driving saccade to pursuit using image motion. *International Journal on Computer Vision* **16**,
- Nordlund, P., Ulihn, T. (1995) Closing the loop: Pursuing a moving object by a moving observer.
- Ornitz, E. M. The development of the vestibulo-ocular reflex from infancy to adulthood. Kaplan, A. R. and Westlake, J. R. *Acta Otolaringologica* 100, 180-193. 1985.
- Panerai, F. and Sandini (1998) Oculo-motor stabilization reflexes: integration of inertial and visual information. *Neural Networks* **11**,
- Panerai, F., Metta, G. and Sandini, G. (2000) Visuo-inertial stabilization in space-variant binocular systems. *Robotics and Autonomous Systems* **30**,
- Rougeaux, S. (1999) Real-time active vision for versatile interaction. Université d'Evry, Courcouronne, France.
- Sharkey, P.M., Murray, D.W., Vandeveld, S., Leid, I.D. and McLauchlan, P.F. (1993) A modular head/eye platform for real-time reactive vision. *Mechatronics* **3**,
- Shelhamer, M., Robinson, D.A. and Tan, H.S. (1992) Context-specific gain switching in the human vestibuloocular reflex. *Ann N Y Acad Sci* **656**, 889-891.
- Shibata, T. and Schaal, S. (1999) Biomimetic Gaze Stabilization. Word Scientific]
- Shupert, C. Development of conjugate human eye movements. *Vision Res* 28, 585-596. 1988.
- Sundaeswaran, V. (1991) Egomotion from Global Flow Field Data. Princeton, NJ - USA
- Ulihn, T., Nordlund, P., Maki, A., Eklundh, J. (1995) Towards an active visual observer.
- Weissman, B. Maturation of the vestibuloocular reflex in normal infants during the first 2 months of life. DiScenna, A. O. and Lisberger SG. *Neurology* 39, 534-538. 1989.
- Wilson, V.J. and Melvill Jones, G. (1988) *Mammalian Vestibular Physiology*, Plenum Press.