Gaze shift reflex in a humanoid active vision system

Ansgar Koene[1,2], Jan Morén[2], Vlad Trifa[2], and Gordon Cheng[2,3]

[1] Knowledge Creating Communication Research Center, NICT, 2-2-2 Hikaridai, Keihanna Science City, Kyoto, 619-0288, Japan,
koene@atr.jp,
WWW: http://www.cns.atr.jp/hrcn
[2] ATR Humanoid Robotics and Computational Neuroscience Laboratories, 2-2-2 Hikaridai, Keihanna Science City, Kyoto, 619-0288, Japan,
[3] JST-ICORP Computational Brain Project, 4-1-8 Honcho, Kawaguchi, Saitama, Japan

Abstract. Full awareness of sensory surroundings requires active attentional and behavioral exploration. In visual animals, visual, auditory and tactile stimuli elicit gaze shifts (head and eye movements) to aid visual perception of stimuli. Such gaze shifts can either be top-down attention driven (e.g. visual search) or they can be reflex movements triggered by unexpected changes in the surroundings. Here we present a model active vision system with focus on multi-sensory integration and the generation of desired gaze shift commands. Our model is being developed based on published data from studies of primate superior colliculus and will be part of the sensory-motor control of the humanoid robot CB.

1 Introduction

Visual perception is often thought of as passive observation. In visual animals however vision is an active process by which we construct a representation of the world from visual inputs in combination with internal and non-visual signals. The human fovea (the only part of the retina with high acuity) covers only about 2° visual angle and yet we effortlessly gain the high spatial resolution information we need to successfully interact with a dynamic environment. This is only possible because, unconsciously, our eyes are continuously scanning, making up to 3-5 saccades (rapid gaze shifts) per second [1]. A more dramatic version of these unconsciously generated gaze shifts are reflex movements towards highly salient, unexpected, auditory, visual or tactile stimuli. These reflex movements can include head and body movements in addition to eye saccades. This gaze shift reflex serves to rapidly reorient the visual sensors in order to the aid information acquisition as required for making a cognitive decision about unexpected stimulus event. In this paper we will present the general architecture for the perceptual system of a humanoid robot featuring multi-sensory (audio-visual) integration, bottom-up salience detection, top-down attentional feature

Proceedings of the ICVS International Cognitive Vision Workshop – ICVW 2007
Published in 2007 by Applied Computer Science Group, Bielefeld University, Germany
This document and other contributions archived and available at: http://biecoll.ub.uni-bielefeld.de

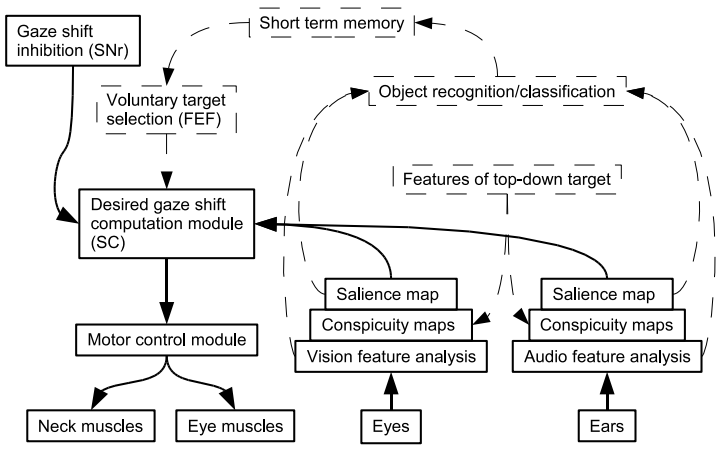

Fig. 1. System Overview: Solid lines indicate implemented modules described in this paper.

gating and reflexive gaze shifting. Following a general introduction to the complete system we will focus on the multi-sensory integration and desired gaze shift computation performed in our "Superior Colliculus (SC)" module.

1.1 Audio-Visual perception system for the Humanoid Robot CB

The sensory-motor system described in this paper comprises the head (& neck) of a 50 degrees of freedom humanoid robot, CB (*Computational Brain*) that was created for exploring the underlying processing of the human brain while dealing with the real world. An overview of the complete system is given in [2]. Here we present only the control of eye & head gaze shifts for audio-visual perception. This will allow the robot to orient its head and eyes so that it can focus its attention on audio and/or visual stimuli. The system will include mechanisms for bottom-up stimulus salience based gaze/attention shifts (where salience is a function of feature contrast) as well as top-down guided search for stimuli that match certain object properties. The main components of the system are shown in figure 1. In order to facilitate interaction with dynamic environments the complete perceptual-motor system must function in real-time. The whole system is therefore being implemented on a distributed computer cluster as described in [2].

Visual feature analysis Following the model proposed by Itti, Koch and Niebur (1998) [3], the visual input is decomposed into several feature streams, each of which calculates feature maps at different scales and combines these into global conspicuity maps, coding the center-surround contrast at each location, for each feature. In order to capture the high salience of sudden stimulus changes a crude form of 'temporal salience' is introduced. A simple low-gain negative

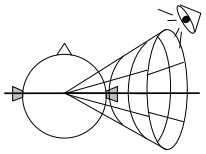

Fig. 2. ITD based sound localization: The head centered cone with symmetry axis running through both ears indicates the 'cone of confusion'. Sound sources on the surface of this cone produce the same ITD.

feedback reduces the conspicuity of stimulus locations where the features have remained unchanged (see [4] for a more detailed study on temporal dynamics). The conspicuity maps for the different visual features are combined into a global saliency map, which encodes the bottom-up saliency of image locations over the entire feature set. The saliency map provides the input to the attention and SC modules driving bottom-up stimulus selection and reflex gaze shifts. The full complement of feature maps is sent to the object recognition/classification module in order to enable feature specific top-down processing.

Auditory feature analysis/localization Auditory stimulus localization is an important component driving attention & gaze shifts, especially when the target is not in sight. Many methods exist for binaural sound localization (see [5] for a review), unfortunately, no method that performs optimally in all natural conditions. The basic principle used by humans and animals to localize sound sources relies on the time difference of arrival of the signals at both ears. The cochlea first performs a frequency band decomposition of the auditory signals. The Inferior Colliculus (ICc) in primates subsequently acts as a temporal correlation detector array that determines the Interaural Time Difference (ITD) between the signals from the left and right ear for each frequency band [6]. This ITD in turn corresponds to a horizontal direction relative to the head. Note, however, that in a bi-aural setup ITD can only restrict the possible sound directions to a conical surface with the symmetry axis running through both ears (figure 2). The ICc output is subsequently used to derive auditory conspicuity maps, indicating for each frequency band where the greatest signal contrast is. These conspicuity maps are then summed over all frequencies (as in ICx [6]) to produce an auditory salience map (figure 3). Again, we increase the relative salience of *changes* in auditory signals by applying a simple low-gain negative feedback that reduces the salience of auditory signal components that have remained unchanged.

Attention Top-down attention to specific feature properties, e.g. guided search, is a distributed process among a number of modules. A feature-gate [7] derived process matches the degree to which the stimulus features (determined by the audio/visual feature analysis processes) at each location match a target feature vector. The degree of stimulus vs. target feature match is propagated to visual and auditory processes where it modulates the saliency maps. The modulation

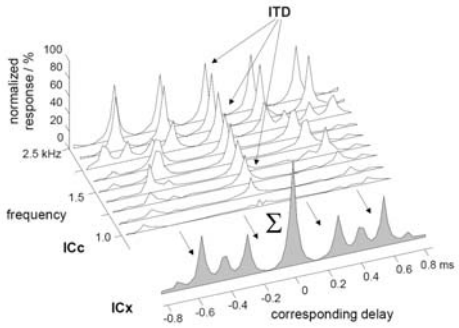

Fig. 3. Auditory salience map: The head centered cone with symmetry axis running through both ears indicates the 'cone of confusion'. Sound sources on the surface of this cone produce the same ITD. (figure reproduced from [6])

strength is set according to the degree of urgency, and the estimated accuracy, of the feature vector representing the search target.

Multi-modal sensory-motor integration (SC module) The Superior Colliculus module combines multi-modal signals with top-down voluntary gaze shift signals and inhibitory gating to produce gaze shift commands. The SC module receives two basic types of inputs, bottom-up excitatory inputs from the sensory processing modules (i.e. vision and audition) and top-down inputs (inhibitory & excitatory) from higher cognitive processing modules. Based on the work by Arai and Keller [8], inhibitory inputs, similar to the inputs from SNr in primates, mediate spatial attention and general sensitivity while excitatory inputs, similar to inputs from FEF in primates, drive deliberate cognitive controlled top-down saccades, e.g. memory saccades. The output of the SC module is a desired gaze shift signal. Decomposition of this signal into eye and head movement components, as well as on-line motor control, is done in subsequent cerebellar and brain stem modules [9].

Object recognition/classification The object recognition module uses the output of the visual and auditory feature maps to performs scene segmentation into probable objects and compares the features of these proto-objects [10] to features of objects that are know to the system. This requires that audio and visual information are coded in the same coordinate system. The salience map is used to prioritize which areas in the scene are analyzed first. The output of this module provides not only the object classification but also the degree of classification certainty. This enables novel, non-recognized, objects to initiate explorative behavior and learning of new objects/classes.

Voluntary target selection The target selection module provides the top-down input to the SC for generating voluntary gaze shifts (i.e. this module is

analogous to FEF in primates [11]). The input to the target selection module is provided by the object recognition/classification module. When searching for a pre-specified target the target selection module first compares the confidently classified objects against the search target. If no matches are found, objects with a classification confidence below a pre-set threshold are considered. The object/area with features most closely fitting the search target feature vector is chosen as gaze shift target. The location of this gaze shift target is sent to the SC module in the form of a sharply tuned excitatory SC activation as well as a more broadly tuned SC inhibition from SNr. The excitatory input provides a sharply tuned drive for a gaze shift to the center of the potential target object while the broadly tuned SC dis-inhibition boost the motor salience of all features in the general area of the potential target object. In the absence of a specified search target the most salient area in perceived space is selected as gaze target. In this case however, salience is a combination of bottom-up feature contrast and top-down novelty (where novelty is inversely proportional to object recognition confidence). The most salient location is promoted as gaze target by reducing the SNr inhibition of SC in that area.

Motor control The motor control module translates desired gaze shift outputs from the SC module into a combination of eye and head movements depending on initial eye and head orientation (eye-head component selection & on-line movement tuning). This module also takes care of the on-line control of the motor execution. In primates this motor control involves the Cerebellum as well as brain-stem circuits that transform desired movement information into rate coded firing patterns sent to motor neurons.

Short term memory The short term memory module stores the location of perceived audio and/or visual objects, keeping track of them when a gaze shift moves these objects out of the field of view, or the audio/visual properties of the object are transient (e.g. a light flash). The short-term memory makes it possible to shift gaze and attention to spatial locations that have in the recent past elicited a perceptual response. In the context of multi-modal bottom-up stimulus driven gaze shifts, short term memory allows for audio-visual interaction between audio and visual inputs that are not simultaneously present. This aids in spatial localization of audio cues by co-localizing them with the most probable visual cue, and helps boost the temporal resolution of visual perception by temporally co-localizing with the most probable audio cue. It is especially important in the context of gaze shifts when a sound comes from an area that was previously looked at but that is not within the current field of view.

2 Multi-modal sensory-motor integration (SC module)

The SC input layer performs a linear weighted summation of the visual and auditory salience maps, the cognitive desired gaze shift activation (FEF) and

the inhibitory spatial attention (SNr) inputs. In order to perform this summation however the inputs must be transformed into a unitary coordinate frame. Since the output of the SC codes desired *change* in gaze direction, the SC codes signals in eye centered gaze coordinates. At the input side the signals are transformed and placed into a single gaze centered map. At the output the spatial population coded SC activity is transformed into signals coding horizontal and vertical desired gaze shift magnitudes that are sent to the corresponding gaze shift generators in the motor control module.

2.1 Coordinate frame transformations

The properties of the SC inputs are:

Audio signal: coded in 1D head centered coordinates, has a spatial resolution of approximately 10° (in humans) distinguishes only horizontal directions (our current sound localization system is purely ITD based) and has a spatial extent of $2 * 180°$ (since ITD based sound localization can not distinguish between sounds from front or from back).

Visual signal: coded in 2.5D eye centered coordinates [horizontal, vertical & binocular), has a spatial resolution of approximately 2° (peripheral vision wide angle camera system) but a spatial extent of only approximately 120°.

Voluntary movements: e.g. memory saccades, appear to be coded in 2D head-centered coordinates [12], with a spatial resolution in the range of 5° for memory saccades [13] and a spatial extent of 360°.

The properties of the SC output are:

Desired gaze shift output: coded in 2D eye centered coordinates with a spatial resolution of 1 to 2° and a spatial extent of 270° (we are limiting ourselves here to eye+head movements without body movement).

An additional coding transformation that takes place at the output of the SC is the transformation from a spatial (map-like) population code to a Cartesian gaze shift vector, which in primates is signaled by a firing rate code [14].

Audio The transformation of the audio salience map from head centered coordinated into the SC coordinate system consists of the following stages:

1. Current eye orientation dependent mapping of the auditory head-centered salience map into eye centered gaze coordinates (figure 4a,b).
2. Expansion from horizontal 1D map to 2D (horizontal & vertical) map (figure 4c).

In an artificial computational system dynamic remapping can simply be achieved by shifting the coordinate axes (index values) of the audio salience map. For biologically constrained neural networks however this poses an interesting challenge that unfortunately falls outside the scope of the current paper. The vertical spread of the 1D horizontal audio map to form a 2D map can be achieved neurally by fanning out of the connection from the audio input to all vertical locations on the output map that have a corresponding horizontal coordinate.

Vision The retinotopic visual inputs to SC are already centered on the gaze direction and therefore do not require eye orientation dependent remapping. The inputs from the two eyes however are slightly shifted relative to each other when the depth of the visual stimuli differs from the fixation depth. Fortunately, the disparity shift between the left and right eye does not affect the SC output, which signals desired gaze shift, when this output is the center of weight of SC activity (and not the location of peak activity as would be the case with a winner-take-all approach). In the case of a version-vergence controlled gaze system, like the primate oculomotor system, gaze direction corresponds to the average direction indicated by both eyes. In the case of stimuli that have binocular disparity a simple projection of the left and right eye information directly onto the SC map, without 'correcting' for vergence, will generate two (usually overlapping) hills of activity in the SC that are left and right of the version direction. Thus, the center of weight of the SC activation still codes the correct desired gaze change. The peak hight however will depend on the degree of overlap between the two hills of activation. In combination with the inhibitory SNr inputs this has the consequence that stimuli at the fixation depth (where the overlap between left and right eye locations is at maximum) can elicit gaze shifts more easily than stimuli at different depths.

Voluntary movement Top down inputs come in two flavors, (i) the excitatory input from FEF that generates a desired gaze shift activation of SC and (ii) inhibitory SNr input that modulates the sensitivity to bottom-up stimulation (globally and locally). Both of these are assumed to be coded in gaze coordinates thus not requiring any coordinate transformation. Microelectrode based recordings of FEF activity in primates during saccades provide evidence that the FEF activity is shifted to correspond to the new eye orientation after saccades [15], which is compatible with the notion that FEF codes information in eye centered coordinates.

Population code to Cartesian vector code transformation This transformation is easily achieved by using differentially weighted connection strengths from the different SC subpopulations to the four groups of brain-stem neurons coding for Up, Down, Left and Rightward movements, where the connection strength increases with desired gaze shift amplitude. In order for this to work the total activity of SC projecting to the output must be constant, independent of the strength of the SC input signals. In order to achieve this the SC input layer does not directly provide the motor output. Rather it projects to an output layer which, through global recursive feedback normalizes total SC activity. When implemented as a recursive neural network this normalization produces activity that looks like the build-up neuron activity in SC [16].

2.2 Computation of desired gaze shift

Based on electrophysiological and behavioral data from primate studies [8], the desired gaze shift is determined by a weighted summation of the excitatory audio,

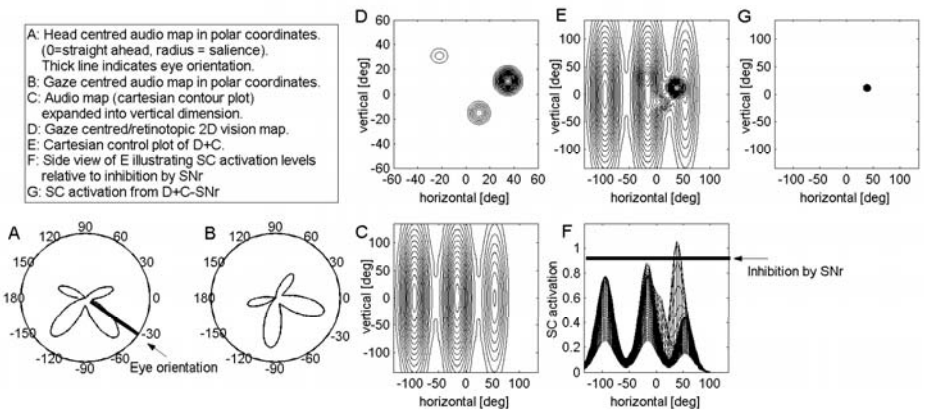

Fig. 4. SC computation of desired gaze shift

visual and voluntary (FEF) inputs with subtraction of the inhibitory SNr input (figure 4). The inhibitory SNr input sets the minimum level of activation that is required for any stimulated areas to contribute to the gaze shift output signal. The center of weight of SC activation determines the desired gaze shift and is automatically determined from the SC population activity by the population to Cartesian vector code transformation at the SC output.

3 Discussion

We have outlined a general framework for an active explorative vision system with emphasis on reflex gaze shift control. Reflex gaze shifts provide humans and animals with a rapid initial response to stimuli that signal potentially important events in their environment. The rapid reorienting of the eyes aids sensory information uptake for making informed decisions. Cognitive vision systems, especially when combined with mobile platforms such as robots, will need to incorporate such a reflex if they are to interact in natural, dynamically changing environments. The core element presented in this paper is the multi-sensory integration module (SC) that generates desired gaze shift commands. This module is closely inspired by the primate Superior Colliculus since the aim of our project is to better understand primate sensory-motor processes through humanoid robotics. In contrast to most of the literature on stimulus driven visual attention (e.g. [17], [18], [19]) the proposed reflex gaze system does not at any stage involve a winner-take-all process. Winner-take-all unambiguously selects the most salient location to become the target of the next gaze shift, guaranteeing that gaze shifts are directed onto a single target. Winner-take-all however has some disadvantages: (1) When implemented in a neural system winner-take-all usually takes the form of an iterative process, which takes time to resolve. This can be a drawback for a rapid response reflex system. (2) No information is given about relative salience in comparison to other areas. (3) In the context

of binocular vision systems, winner-take-all can not automatically extract the version information from the combination of the left and right eye images, instead the left and right eye images must be averaged before the winner-take-all stage. Instead we use inhibitory (SNr) input to control the required minimum level of stimulus salience to elicit a gaze shift response. Resulting eye movements are based on the center of weight of the remaining SC activity. In the context of reflex gaze shifts this inhibition input will usually be so high that only very salient events, such as sudden movement, a bright flash or a loud bang, will be strong enough to elicit a gaze shift response. An obvious drawback is that in the rare event where multiple highly salient stimuli are present at the same time, the center of weight of SC activity may lie in-between the salient locations eliciting an incorrect movement. Since reflex triggering events/stimuli are low-probability events/stimuli, this is not likely to happen. Furthermore, as was shown by Arai and Keller [8], primate saccades under conditions with multiple targets validate this behavior. Unlike winner-take-all, finding the center of weight of SC activity does not add any computational cost to a real-time reflex response system since the center of weight computation is a by-product of the spatial to Cartesian code transformation which is required since the brain stem saccade generator system uses Cartesian coded information [14]. Natural visual scene scanning behavior should also be easily achievable. Start with a relatively low level of SNr inhibitory input to all areas of SC. Shift the gaze according to the center of weight of SC activity. Once an area has been visually scrutinized, increase the SNr inhibition input for this area and move to the new center of weight of SC activity, thus gradually, piecewise, increasing the SNr inhibition of all areas. In order to track which areas have already been scrutinized the short term memory module has to track and re-map viewed areas as a function of current gaze direction.

The gaze shift reflex proposed in this paper provides merely the stimulus driven foundation of the active vision system. When considering the challenges of integrating this reflex with top-down voluntary gaze shift control one of the first issues to consider will be how to set and adjust the reflex sensitivity (level of inhibitory input from SNr). Ultimately this should be context dependent [20] and thus driven by the object recognition module, or a scene recognition module that combines the object recognition and localization information to interpret the surroundings. For the object recognition system it will be vital to enable the humanoid robot to perform exploratory movements and manipulation of its surroundings so that it can acquire object recognition feature-vectors.

References

1. L. Lunenburger, W.L., Hoffmann, K.P.: Neural activity in the primate superior colliculus and saccadic reaction times in double-step experiments. Progress in Brain Research **142** (2003) 91–107
2. Cheng, G., Hyon, S.H., Morimoto, J., Ude, A., Jacobsen, S.C.: Cb: A humanoid research platform for exploring neuroscience. In: IEEE-RAS/RSJ International Conference on Humanoid Robots (Humanoids 2006). (2006)

3. L. Itti, C.K., Niebur, E.: A model of saliency-based visual attention for rapid scene analysis. IEEE Trans. Patt. Anal. Mach. Intell. **20** (1998) 1254–1259
4. J. Triesch, D.H.B., Jacobs, R.A.: Fast temporal dynamics of visual cue integration. Perception **32** (2002) 421–434
5. J. Chen, J.B., Huang, Y.A.: Time delay estimation in room acoustic environments: an overview. EURASIP Journal on applied signal processing **2006** (2006) 1–19
6. Schauer, C.: Modellierung primärer multisensorischer Mechanismen der räumlichen Wahrnehmung. PhD thesis, Fakultät für Informatik und Automatisierung der Technischen Universität Ilmenau (2006)
7. Cave, K.R.: The featuregate model of visual selection. Psychological Research **62** (1999) 182–194
8. Arai, K., Keller, E.L.: A model of the saccade-generating system that accounts for trajectory variations produced by competing visual stimuli. Biological Cybernetics **92** (2005) 21–37
9. Quinet, J., Goffart, L.: Saccade dysmetria in head-unrestrained gaze shifts after muscimol inactivation of the caudal fastigial nucleus in the monkey. Journal of Neurophysiology **93**(4) (2005) 2343–2349
10. Rensink, R., Enns, J.: Early completion of occluded objects. Vision Research **38** (1998) 2489–2505
11. Schiller, P.H., Tehovnik, E.J.: Neural mechanisms underlying target selection with saccadic eye movements. Progress in Brain Research **149** (2005) 157–171
12. Niemeier, M., Karnath, H.: Stimulus-driven and voluntary saccades are coded in different coordinate systems. Current Biology **13**(7) (2003) 585–598
13. Ohtsuka, K., Sawa, M.: Accuracy of memory-guided saccades. Ophthalmologica **198** (1989) 53–56
14. Groh, J.M.: Converting neural signals from place codes to rate codes. Biological Cybernetics **85**(3) (2001) 159–165
15. Nakamura, K., Colby, C.: Updating of the visual representation in monkey striate and extrastriate cortex during saccades. Poc. Natl. Acad. Sci. U.S.A. **99**(6) (2002) 4026–4031
16. S. Everling, M.C. Dorris, R.K., Munoz, D.: Role of primate superior colliculus in preperation and execution of anti-saccades and pro-saccades. Journal of Neuroscience **19**(7) (1999) 2740–2754
17. D.K. Lee, L. Itti, C.K., Braun, J.: Attention activates winner-take-all competition among visual filters. Nature Neuroscience **2** (1999) 375–381
18. Koch, C., Ullman, S.: Shifts in selective visual attention: towards the underlying neural circuitry. Human Neurobiology **4** (1985) 219–227
19. Zhaoping, L.: A saliency map in primary visual cortex. TRENDS in Cognitive Sciences **6**(1) (2002) 9–16
20. C.Cheng, A.N., Kuniyoshi, Y.: Continuous humanoid interaction: An integrated perspective - gaining adaptivity, redundancy, flexibility - in one. Robotics and Autonomous Systems **37**(2-3) (2001) 161–183