Humanoid Vision Resembles Primate Archetype

Andrew Dankers[1,3], Nick Barnes[2,3], Walter F. Bischof[4], and Alexander Zelinsky[5]

[1] Italian Institute of Technology, Via Morego 30, Genova Italy 16142
andrew.dankers@iit.it
[2] National ICT Australia[4], Locked Bag 8001, Canberra ACT Australia 2601
[3] Australian National University, Acton ACT Australia 2601
nick.barnes@nicta.com.au
[4] University of Alberta, Edmonton, Alberta Canada T6G2E8 wfb@ualberta.ca
[5] CSIRO ICT Centre, Canberra ACT Australia 0200 alex.zelinsky@csiro.au

Summary. Perception in the visual cortex and dorsal stream of the primate brain includes important visual competencies, such as: a consistent representation of visual space despite eye movement; egocentric spatial perception; attentional gaze deployment; and, coordinated stereo fixation upon dynamic objects. These competencies have emerged commensurate with observation of the real world, and constitute a vision system that is optimised, in some sense, for perception and interaction. We present a robotic vision system that incorporates these competencies. We hypothesise that similarities between the underlying robotic system model and that of the primate vision system will elicit accordingly similar gaze behaviours. Psychophysical trials were conducted to record human gaze behaviour when free-viewing a reproducible, dynamic, 3D scene. Identical trials were conducted with the robotic system. A statistical comparison of robotic and human gaze behaviour has shown that the two are remarkably similar. Enabling a humanoid to mimic the optimised gaze strategies of humans may be a significant step towards facilitating human-like perception.

1 Introduction

Biologically-inspired active vision mechanisms exhibiting primate-like agility (e.g., *CeDAR* [20], and *iCub* [18]; Fig.1) permit the investigation of primate-like visual competencies. Primates have evolved invaluable visual abilities which provide a level of perception that enables intelligent cognition. These abilities include foveal vision and gaze strategies that facilitate efficient perception, such as the propensity to attend locations containing relevant visual information. They constitute the basic visual abilities we wish to synthesise in the development of artificial cognitive systems that operate in the real world.

[4] National ICT Australia is funded by the Australian Department of Communications, Information Technology and the Arts and the Australian Research Council through *Backing Australia's ability* and the ICT Centre of Excellence Program.

Though components of the robotic vision system take biological inspiration, we focus on the development of a system that reproduces the visual behaviours of its primate archetype by incorporating similar competencies, rather than by developing an exacting reconstruction of the underlying processes in the primate brain. We hypothesise that similarities between the underlying robotic system model and that of the primate vision system will elicit similar gaze behaviours. Accordingly, psychophysical trials were conducted to record human gaze behaviour when free-viewing a reproducible, dynamic, 3D scene. Identical trials were conducted with the robotic system. A statistical comparison of the robotic and human gaze behaviour was then conducted.

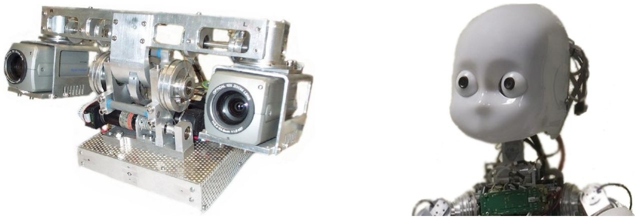

Fig. 1. *CeDAR* (left); and *iCub* (right).

2 System Archetecture

Primate-inspired components of the robotic vision system 2 include spatiotemporal registration of camera images into a rectified egocentric reference frame (Section 2.1), 3D space-variant spatiotemporal representation of visual surfaces (Section 2.2), coordinated foveal fixation upon, and tracking of, attended surfaces (Section 2.3), and a novel attention system (Section 2.4). The processing components are portable to active vision systems such as the iCub and CeDAR mechanisms. Moreover, the core software is available under open-source release in collaboration with the *RobotCub*[6] project.

2.1 Egocentric Perception

Humans experience spatiotemporal continuity when integrating actively acquired imagery into a unified perception. Mechanisms of spatial updating maintain accurate representations of visual space across eye movements. Furthermore, binocular imagery is combined into a singular egocentric representation that accounts for gaze convergence. Monkeys too retain consistent representations of visual space across eye movements by transferring activity among spatially-tuned neurons within the intraparietal sulcus [12].

For a robotic active stereo system, camera pan and tilt motions introduce image perspective distortions. Barrel distortions may additionally be introduced by camera lenses. Synonymous with kinesthetic feedback from ocular muscles in the primate eye, online evaluation of epipolar geometry from encoder data is used to account for the image-frame effect of gaze convergence,

[6] www.robotcub.org

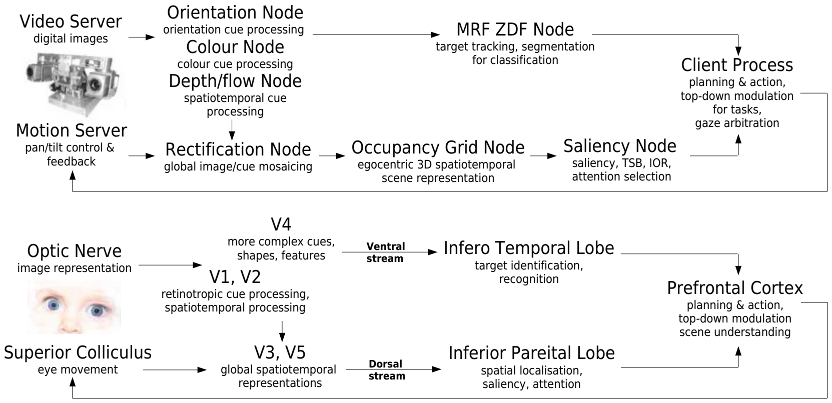
Video Server
digital images
Orientation Node
orientation cue processing
Colour Node
colour cue processing
Depth/flow Node
spatiotemporal cue processing
MRF ZDF Node
target tracking, segmentation for classification
Client Process
planning & action, top-down modulation for tasks, gaze arbitration
Motion Server
pan/tilt control & feedback
Rectification Node
global image/cue mosaicing
Occupancy Grid Node
egocentric 3D spatiotemporal scene representation
Saliency Node
saliency, TSB, IOR, attention selection
V4
more complex cues, shapes, features
Ventral stream
Infero Temporal Lobe
target identification, recognition
Optic Nerve
image representation
V1, V2
retinotropic cue processing, spatiotemporal processing
Prefrontal Cortex
planning & action, top-down modulation scene understanding
Superior Colliculus
eye movement
V3, V5
global spatiotemporal representations
Dorsal stream
Inferior Pareital Lobe
spatial localisation, saliency, attention

Fig. 2. Block diagram showing major feedforward data flow between functional nodes in the robotic vision system (top); and, a summary of major feedforward interactions between functional regions in the primate visual brain (bottom).

facilitating the registration of imagery into an egocentrically static reference frame across camera pan and tilt motion. We can project camera images into this reference frame, and from this reference frame to one that spatiotemporally corresponds to the real world and other sensing modalities, such as an egosphere or occupancy grid (Fig.3, Section 2.2). In [1], we described a method to rectify camera barrel distortions and to register images in mosiacs exhibiting global *parallel epipolar geometry* [6]. Moreover, online epipolar rectification of camera imagery, and the projection of such rectified images into globally fronto-parallel rectified mosaics enables the use of *static* stereo algorithms, such as those that depend on fronto-parallel geometry, on *active* stereo platforms.

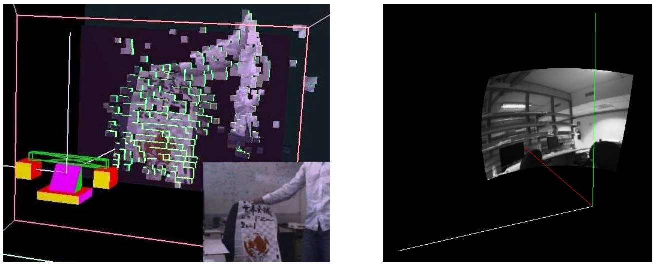

Fig. 3. Realtime egocentric 3D scene reconstruction (left, inset left camera view), and projection of imagery into an egosphere (right).

2.2 Spatiotemporal Perception

Recent investigations into primate spatial perception suggest a separation of the estimation of relative retinal disparity from the conversion to absolute scene depths [15]. Other research provides evidence suggesting that processing of retinotropic and absolute motion occurs in separate areas in the primate brain [17, 9]. The representation of visual space matures from retinotropic in early life to egocentric, coinciding with the development of specific cortical areas [8, 5]. Gaze convergence, focal length and prior familiarity with an object's size can provide information for conversion from relative to absolute depth distances. Gaze convergence stretches extraocular muscles, from which kinesthetic sensations project to the visual cortex where they facilitate absolute depth perception [22].

Synonymous with primate occular kinesthetic feedback, images are registered within the epipolar rectified mosaics based upon encoder data. This converts relative disparity estimation in the image frame to a (1D) search for absolute disparities in the static mosaic reference frame. Absolute disparity estimations are integrated into a space-variant Bayesian occupancy grid (left, Fig.3) tailored for use with stereo vision sensing, in realtime. 2D optical flow is also estimated in mosaic space, which removes the image-frame effect of deliberate camera motion. Re-projection of the camera images, or cues extracted from camera images, onto the occupancy grid establishes cue-surface correspondences. In this manner, a representation of the location of visual surfaces in the scene, their coarse 3D structure and motion, and their appearance and cue responses, is obtained.

2.3 Coordinated Fixation & Target Segmentation

Monkeys exhibit vigorous neuronal responses when viewing small laboratory stimuli in isolation, compared to the sparse neuronal activity elicited when viewing broad scenes [21]. Long range excitatory connections in V1 appear to enhance responses of orientation selective neurons when stimuli extend to form a contour [4]. During binocular fixation, the foveas align over an attended target in a coordinated manner. An attended object appears at near identical left and right retinal positions, whereas the rest of the scene usually does not; that is, the attended object exhibits *zero disparity*.

Various synthetic targeting systems use correlation methods, or extract 'blobs' from images to track a target, and typically select a target location for the left and right cameras independently. Perspective distortions and directional illumination effects, amongst other causes, may yield left and right camera fixation points that do not accurately correspond to the same real scene point. Rather, coordinated primate-like stereo fixation incorporating rapid, model-free target tracking and accurate foveal target segmentation is achieved using a robust *Markov random field zero disparity filter* (MRF ZDF) [2]. The formulation uses stereo image data to enforce optimal retinal alignment of the centre of the left and right cameras with a selected scene location, regardless of its appearance and foreground or background clutter, without relying upon independent left and right target extraction.

2.4 Attention

Navalpakkham *et al.* [14], amongst others, suggest that because neurons involved in attention are found in different parts of the brain that specialise in different functions, they may encode different types of visual salience: they propose that the posterior parietal cortex encodes visual salience; the prefrontal cortex encodes top-down task relevance; and the final eye movements are subsequently generated in the superior colliculus where attentional information from both regions is integrated. In accordance with this proposal, we compute an *attention mosaic* as the product of three intermediary maps: a retinotopic saliency map, an active-dynamic *inhibition of return* (IOR) map, and a *task-dependent spatial bias* (TSB) map. Finally, covert moderation of peaks in the attention mosaic filters the selection of the next scene point that will receive overt attentional fixation.

Visual Saliency: We adopt the widely accepted bottom-up model of attention [7] extended specifically for active cameras and dynamic scenes. Approximations of the retinal ganglion center-surround response is computed to determined uniqueness in various cue maps including intensity, intensity-normalised colour chrominance, colour distance, depth and optical flow. From a log-Gabor[7] phase analysis, orientation saliency, symmetry, and phase-congruent corner and edge cue maps are obtained [11]. For each cue map, a *difference-of-Gaussian* (DOG) image pyramid approach provides multi-scale center-surround responses. Saliency cues are combined into a single saliency map.

Inhibition of Return: Primates transiently inhibit the activity of neurons associated with the saliency of an attended location [10]. Further, in the intraparietal sulcus of monkeys, the activity of spatially-tuned neurons corresponding to the location of a salient stimulus was shown to be transferred to other neurons commensurate with eye motion [12], a concept known as *efference copy* that assists prediction of the position of the eyes (and other body parts). A Gaussian inhibition kernel is added to the region around the current fixation point in an IOR accumulation mosaic, and decayed, every frame. Expanding upon this for dynamic scenes, accumulated IOR is propagated in egocentric mosaic space according to optical flow. In this manner, IOR accumulates at attended scene locations, but it remains attached to objects as they move. Propagated IOR is spread and reduced according to positional uncertainty. We decrement IOR over time so that previously inhibited locations eventually become uninhibited.

Task-Dependent Spatial Bias: The prefrontal cortex implements attentional control by amplifying task-relevant information relative to distracting stimuli [16]. We introduce a TSB mosaic that can be dynamically tailored according to tasks. TSB can be preempted for regions not in the current view frame but within the broader mosaic.

Attention & Saccade Moderation: An image-frame attention map is constructed as the product of the saliency, IOR and TSB maps. Attention map

[7] log-Gabor responses have been observed in orientation sensitive neurons in cats [19] and exhibit a broader spatial response than traditional Gabors.

peaks are covertly moderated before the overt fixation point is selected. Several types of moderation have been implemented: *supersaliency* - a view frame coordinate immediately wins attention if it is significantly more salient than the next highest peak in the attention map; *clustered saliency* - attention is won by the view frame location about which numerous global peaks occur within several consecutive frames. If neither of the above winners emerge sufficiently rapidly, attention is given to the highest peak in the attention map since the previous fixation location was selected.

3 Psychophysical Trials

Similarities between the underlying humanoid system model and that of the primate vision system are hypothesised to elicit respectively similar basic gaze behaviours. Accordingly, 20 human and 4 robotic trials were conducted where 3D visual stimuli were moved in a reproducible manner within a bounded scene volume (Fig. 4). A non-intrusive gaze tracker recorded the path of the human participants' gaze (left, Fig.5). Identical trials were conducted with the robotic system for statistical comparison to commonalities found in the human trial data.

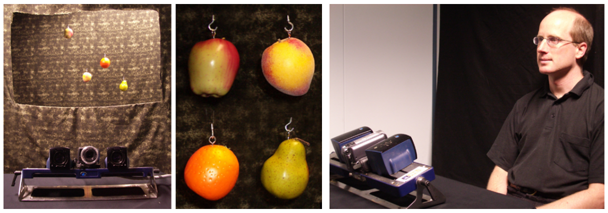

Fig. 4. Psychophysical trials: participant's view (left); trial stimuli (centre); non-intrusive gaze tracking (right).

3.1 Human Benchmark Trials

Two pilot trials were initially conducted to observe emergent human gaze behaviours, and to determine how such behaviours could be characterised statistically. Histograms of gaze velocity magnitude data (right, Fig.5) from the human trials exhibited a distinctly bimodal appearance - much of the gaze path was attended at either near zero (smooth pursuit, or tracking) velocities, or high (saccade, or attentional shift) velocities, with few frames exhibiting medial velocities. For each trial, a threshold was selected within the medial velocity range above which the elicited inter-frame gaze velocity magnitudes were labeled as saccades, and below which they were considered smooth pursuit (centre, Fig.5). Each data point was also marked according to whether it was recorded during a period when a scene object was translating (T periods), or when no objects were translating (NT periods). Histograms and spatial plots of gaze velocity and position data during only T, and during only NT were also constructed.

Based upon empirical observations, 13 trial parameters (a non-limiting set) were extracted from each human trial data log (left, Table 1). To reduce the impact of participant mood/alertness, ratio parameters between T and NT were extracted from each trial providing seven pseudo-normalised statistics suitable for inter-individual comparison (right, Table 1). For the object re-attention period parameter, the standard deviation of object re-attention periods for each object in a trial was used as a pseudo-normalised metric to estimate coherence to a constant object re-attention period over a trial: $P_{sd} = \mathrm{STD}(P_o)$, (where $o = 0...4$, corresponding to separate re-attendance periods P_o for each of the four separate objects presented during each trial).

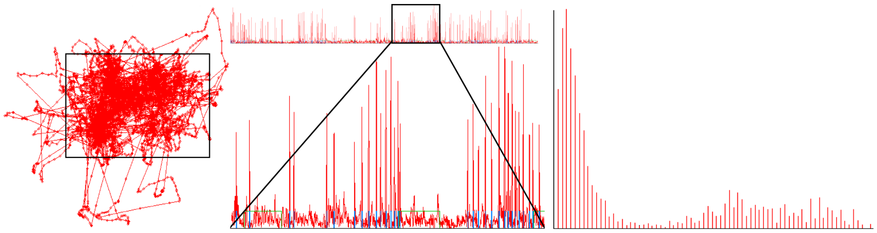

Fig. 5. Data for a single human trial (units ommited): 2D projection of complete gaze path with location of scene window (left); gaze velocity magnitude time-line (centre, above) with enlargement (centre, below) showing saccades (blue) and periods of object translation (green); and, histogram of velocity magnitudes (right).

The small sample size (20 trials) makes it difficult to confirm that the underlying probability distribution functions (PDFs) associated with the extracted rate parameters conform to normal distributions. For example, both JB and KS tests for PDF normality [13] fail for most rate parameters unless less restrictive thresholds are chosen than recommended. Consequently, we *bootstrap*[3] the distribution of means and variances for each rate parameter. The red bars in Fig.6 summarise the bootstrapped 95% confidence intervals (CIs) on the mean and standard deviations for each inter-individual rate parameter, calculated over all data from all human trials. The plotted bootstrapped intervals indicate whether the inter-individual rate parameter is characteristically likely to increase or decrease when transitioning from T to NT, according to its location above or below 1.0 (respectively). The last parameter, the re-attention period coherence parameter (P_{sd}), is an absolute measure obtained during NT in each trial.

3.2 Robotic Trials

Robotic trials were then conducted using the same trial apparatus and stimuli as for the human participants. After each trial, configuration settings were iteratively adjusted such that the system was deemed likely to elicit behaviours more similar to human performance. Ratio parameters, and the re-attention consistency parameter, were extracted from each robotic trial for comparison with the human rate parameter behavioural statistics.

Table 1. Extracted average absolute trial parameters (left), and parameters used for inter-individual behavioural statistics (right).

Spt_t, Spt_{nt}	smooth pursuit durations	$Spt_r = Spt_{nt}/Spt_t$
Spl_t, Spl_{nt}	smooth pursuit distances	$Spl_r = Spl_{nt}/Spl_t$
Spv_t, Spv_{nt}	smooth pursuit velocities	$Spv_r = Spv_{nt}/Spv_t$
Scl_t, Scl_{nt}	saccade distances	$Scl_r = Scl_{nt}/Scl_t$
Scv_t, Scv_{nt}	saccade velocities	$Scv_r = Scv_{nt}/Scv_t$
Scf_t, Scf_{nt}	saccade frequency	$Scf_r = Scf_{nt}/Scf_t$
P	object re-attention period during NT	$P_{sd} = \text{STD}(P_o)$

Subscripts denote measurement period - t: translation, nt: no translation.

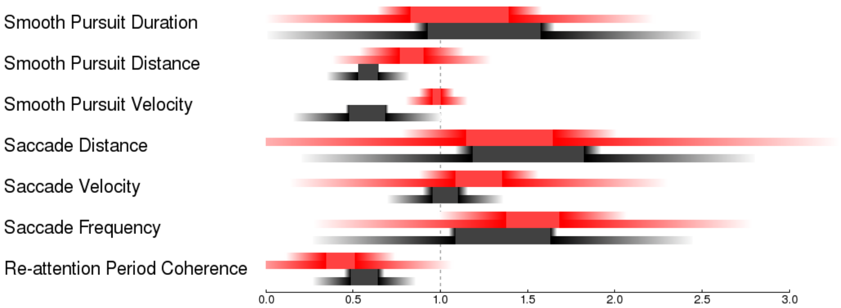

Fig. 6. Bootstrapped human (red) and robotic (black) inter-individual rate parameters. Distributions represent the rate change from periods where an object is translating (T) to periods where no objects are translating (NT). Each solid central bar region represents the bootstrapped 95% CI for the distribution of means, calculated from all average rate parameters extracted from all trials. Upper and lower fading bars represent the 95% CI lower and upper bounds (respectively) of *two* bootstrapped standard deviations. Significant correlation exists between the human and robotic rate parameter distributions.

4 Statistical Comparison

It is often possible to compare the performance of a system to a theoretical model by monitoring output and performing model-based residual analyses. However, primate gaze behaviours are the product of a complex biological system. There is no general theory of human gaze behaviour that would permit such a systematic comparison. It is nevertheless possible to conduct a 'black-box' comparison of the gaze behaviours of humans and machines by comparing the statistics and PDFs associated with specific parameters derived from output gaze behaviours elicited by common input stimuli. In this regard, cluster overlap and KL divergence methods [13] to compare gaze parameters may not be appropriate due to small sample sizes in the human (20 samples) and robotic (four non-independent samples) trials. Therefore, the bootstrapped human statistics are used as a set of benchmarks to which the same parameters extracted individually from each robotic trial are compared. Accordingly, each rate parameter in each robotic trial was examined to determine if it fell within one, and then two bootstrapped standard deviations of the corresponding bootstrapped human inter-individual parameter means. The majority of extracted robotic parameters fell within one 95% CI bootstrapped upper-bound standard deviation of the corresponding human benchmark. All but parameter Spv_r fell within two bootstrapped 95% CI standard deviations of the upper bound of the bootstrapped 95% CI mean. This single discrepancy is likely due to the low accuracy (low signal to noise ratio) involved in detecting small, low velocity eye motions with FaceLAB.

As methodologically expected, robotic trial 4 performed the best in terms of extracted parameters best conforming to human benchmark statistics. Nevertheless, *all* trials exhibited good conformity to the bootstrapped human statistics. Moreover, the system was observed to produce human-like behaviours in all trials, regardless of the wide variance in configuration settings. This suggests the behaviours elicited are largely dependent on the implemented system model, not just the configuration settings selected for a particular trial. As a case in point, if considered as a set of four independent samples, the robotic group statistics may be bootstrapped for comparison to the bootstrapped human group statistics. The black bars in Fig.6 show that when considering all robotic trials as independent samples of a single underlying PDF, the bootstrapped robotic mean rates consistently change in the same direction as the bootstrapped human rates: where human rates tended to increase in going from T to NT, so did the robotic rates. Of course, the robotic trials were *not* conducted completely independently, so this is not a strong claim. It is however noted that there is considerable overlap between the bootstrapped human and robotic group parameter statistics in Fig.6.

5 Conclusion

Even though system components take biological inspiration, the trials do not provide information about the *structural* similarity of the system, or its components, to the primate visual brain. They may only be used to compare benchmarks obtained from human trials with the emergent gaze behaviours of a robotic system which incorporates primate-inspired competencies. The fact that all robotic trials, all with different configuration settings, exhibited a majority of behavioural parameters that fell within the bootstrapped standard deviations of human benchmark behavioural parameters, suggests that the behaviour of the robotic system is largely a product of the underlying biologically-inspired model. Though the assumption that all trials may be treated as individual sample points is weak, when treated as such, the group statistics thus formed also conform well to the human benchmarks. Nevertheless, the strong conformity of individual robotic trial behavioural parameters to the corresponding human benchmarks indicates that, in terms of these trials, the primate-inspired humanoid system achieves primate-like gaze behaviours, for this task.

References

1. A. Dankers, N. Barnes, and A. Zelinsky, “Active vision - rectification and depth mapping,” in *Aust. Conf. Robotics and Automation*, 2004.
2. ——, “Mapzdf segmentation and tracking using active stereo vision: Hand tracking case study,” in *Comp. Vis and Im. Understanding*, 2006.
3. B. Efron and R. J. Tibshirani, “An introduction to the bootstrap,” in *Chapman and Hall*, 1993.
4. C. Gilbert, M. ito, M. Kapadia, and G. Westheimer, “Interactions between attention, context and learning in primary visual cortex,” in *Vision Res.*, 2000.
5. R. O. Gilmore and M. H. Johnson, “Learning what is where: Oculomotor contributions to the development of spatial cognition,” in *The Development of Sensory, Motor and Cognitive Capacities in Early Infancy (pp. 25-47)*, 1998.
6. R. Hartley and A. Zisserman, *Multiple View Geometry in Computer Vision, Second Edition.* Cambridge University Press, 2004.
7. L. Itti, “Models of bottom-up attention and saliency,” in *Neurobiology of Attention*, 2005.
8. M. H. Johnson, “Developmental cognitive neuroscience, vol. 1, 3 ed. malden,” in *MA and Oxford UK: Balckwell Publisher Inc.*, 1997.
9. E. R. Kandel, J. H. Schwartz, and T. M. Jessell, “Principles of neural science, 4th edition,” in *McGraw-Hill Medical*, 2000.
10. R. M. Klein, “Inhibition of return,” in *Trends in Cognitive Sciences*, 2000.
11. P. Kovesi, “Phase congruency detects corners and edges,” in *Aust. Patt. Rec. Soc.*, 2003, pp. 309–318.
12. E. Merriam, C. Genovese, and C. Colby, “Spatial updating in human parietal cortex,” in *Neuron*, 2003, pp. 39:351–373.
13. T. M. Mitchell, “Machine learning,” in *McGraw-Hill*, 1997.
14. V. Navalpakkam, M. Arbib, and L. Itti, “Attention and scene understanding,” in *Neurobiology of Attention*, 2005.
15. P. Neri, H. Bridge, and D. J. Heege, “Stereoscopic processing of absolute and relative disparity in human visual cortex,” in *Neurophysiol. 92: 18801891*, 2004.
16. S. Nieuwenhuis and N. Yeung, “Neural mechanisms of attention and control: losing our inhibitions?” in *Nature*, 2005, pp. 8:1631–1633.
17. S. Nishida, T. Ledgeway, and M. Edwards, “Dual multiple-scale processing for motion in the human visual system,” in *Vision Research 37: 2685-2698*, 2001.
18. G. Sandini, G. Metta, and D. Vernon, “The icub cognitive humanoid robot: An open-system research platform for enactive cognition,” in *50 Years of AI, Springer-Verlag pp. 359370*, 2007.
19. Sun and Bonds, “Two-dimensional receptive field organization in striate cortical neurons of the cat,” in *Vis Neurosci.*, 1994, pp. 11: 703–720.
20. H. Troung, “Active vision head,” in *Thesis, Australian Nat Univ.*, 1998.
21. W. E. Vinje and J. L. Gallant, “Sparse coding and decorrelation in primary visual cortex during natural vision,” in *Science 287:12731276*, 2000.
22. J. L. Zajac, “Convergence, accommodation, and visual angle as factors in perception of size and distance,” in *American Journal of Psychology, Vol. 73, No. 1*, 1960.