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* [30], and *iCub* [26]; Fig.1) permit the investigation of primate-like visual competencies. Vision is a data-rich sensing modality useful for environmental perception, navigation, search, hazard and novelty detection, and communication. Primates have evolved invaluable visual abilities which

[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.

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.

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. This is partially a consequence of the fact that the function of the visual cortex is not precisely known, and because the hardware with which it is synthesised differs from that of the visual cortex. Nevertheless, 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.

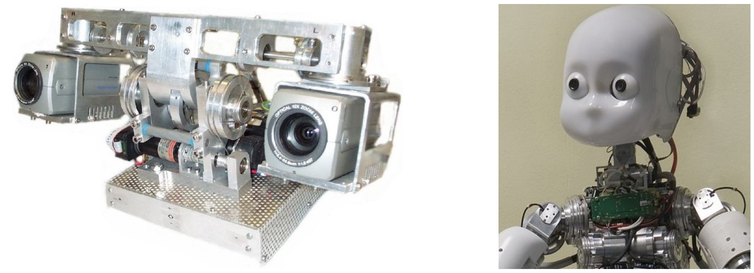

Fig. 1. *CeDAR* (left); and *iCub* (right).

2 System Archetecture

Components of the realtime robotic vision system include spatiotemporal registration of camera images into a rectified egocentric reference frame (Section 2.1), a 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 exhibiting high apparent resolution.

[6] www.robotcub.org

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. Long straight lines are indeed perceived as straight and continuous in our cyclopean perception, also if they exist in the visual fields of both eyes. Monkeys too retain consistent representations of visual space across eye movements by transferring activity among spatially-tuned neurons within the intraparietal sulcus [19].

For a robotic active stereo system, camera pan and tilt motions introduce image perspective distortions. Barrel distortions may additionally be introduced by camera lenses, yielding images in which straight edges appear curved. For spatiotemporal registration and left-right integration of active stereo imagery into a unified, head-centered, human-like perception, such phenomena must be accounted for. 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 facilitating the registration of imagery into a unified perception across camera pan and tilt motion. Any curvature in the image-frame projection of straight lines can be removed so that the lines appear straight and continuous across binocular imagery. The pivot in egocentric perception is to register images in an egocentrically static reference frame. We can project camera images into this reference frame, and vice versa, 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.2, Section 2.2), and vice versa. In [2], we described a method to rectify camera barrel distortions and to register images in mosiacs exhibiting global *parallel epipolar geometry* [10]. 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. Estimates of stereo disparity, for example, can be used for spatial perception in the vicinity of the attended scene location. In this manner we achieve a coarse, probabilistic, realtime, egocentric 3D Bayesian occupancy grid reconstruction of scene structure and motion in the vicinity of the gaze fixation location[7]. Images (and processed cue responses) can be re-projected onto the internal 3D scene representation, enabling a realtime perception of the location, motion and appearance of visual surfaces.

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 [22]. Other research provides evidence suggesting that processing of retinotropic and absolute motion occurs in separate areas in the primate

[7] Stereo *fixation* involves the alignment of the optical centres of the left and right cameras with a specific scene point.

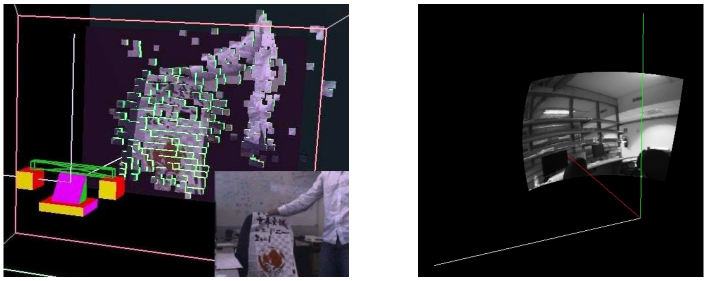

Fig. 2. Realtime egocentric 3D scene reconstruction (left, inset shows left CeDAR camera view), and projection of imagery into an egosphere that spatiotemporally corresponds to the real world (right, showing head-centered coordinate system – red vector shows direction iCub nose points).

brain [24, 16]. The representation of visual space matures from retinotropic in early life to egocentric, coinciding with the development of specific cortical areas [14, 9] . 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 [32].

Numerous methods exist to calculate relative depth and optical flow within 2D camera projections of a scene. Few such methods calculate absolute scene depth and flow accounting for the image frame effects of deliberate camera egomotion. Synonymous with kinesthetic feedback in the ocular system of primates, images registered within epipolar rectified mosaics using encoder data converts relative disparity estimation in the image frame to a 1D search along horizontal scan-lines for absolute disparities in the static mosaic reference frame. Conducting a disparity search over ±16 pixels in the overlapping region of the currently augmented section of the left and right mosaics defines a measurable scene volume in which structure can be coarsely assessed. Absolute disparity estimations are integrated into a space-variant Bayesian occupancy grid (left, Fig.2) tailored for use with stereo vision sensing, in real-time. The 3D velocities of visual surfaces in the depth direction are calculated using an approach similar to that of [15]. 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 structure and motion, and their appearance and cue responses, can be 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 [31]. Long range excitatory connections in V1 appear to enhance responses of orientation selective neurons when stimuli extend to form a contour [8]. Attention binds the visual attributes of an attended object, such as colour, form and/or texture, into a unitary percept [29, 25]. 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) [4]. 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. Relaxation of the zero disparity constraint facilitates segmentation of dominant retinally aligned surfaces. In this manner, the notion of an object can be defined as a spatially coherent visual surface that projects to the same left and right image coordinates.

2.4 Attention

Robotic target *selection* also takes primate inspiration. Navalpakkham *et al.* [21], 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*[8] of peaks in the attention mosaic filters the selection of the next scene point that will receive overt attentional fixation.

[8] *Covert* moderation involves consideration of peripheral image locations without moving the cameras.

Visual Saliency

Saliency is determined each frame, largely as per the widely accepted bottom-up model of attention [13] extended specifically for active cameras, dynamic scenes, and top-down modulation. A *difference-of-Gaussian* (DOG) approximation of the retinal ganglion center-surround response is adopted to determine uniqueness in various cue maps including intensity, intensity-normalised colour chrominance, colour distance[9], depth and flow. Log-Gabor wavelets are adopted for computational efficiency in conducting an image phase analysis as such wavelets exhibit a broader spatial response[10] than traditional Gabors. From this log-Gabor phase analysis, orientation saliency, symmetry, and phase-congruent corner and edge maps are obtained [18]. For each cue map, an image pyramid approach provides scale-independent center-surround responses. Saliency cues are weighted and added into a single saliency map for each camera. Active rectification and absolute mosaic disparity can be used to combine left and right saliency maps into a single egocentric saliency mosaic.

Inhibition of Return

Primates transiently inhibit the activity of neurons associated with the saliency of an attended location [17]. 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 [19], 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, 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 according to decay rate I_d, so that previously inhibited locations eventually become uninhibited. Faster I_d decay elicits more frequent saccades. This rate can be modulated by higher level client processes.

Task-Dependent Spatial Bias

The prefrontal cortex implements attentional control by amplifying task-relevant information relative to distracting stimuli [23]. We introduce a TSB mosaic that can be dynamically tailored according to tasks. For example, when driving a car, humans tend to keep their eyes on the road - we may synthesise this tendency by biasing the lower half of the mosaic where we may

[9] The Malhonobis distance from any selected target chrominances

[10] log-Gabor wavelets are similar to the impulse response observed in the orientation sensitive neurons in cats [28].

expect to find the road, regardless of the current gaze location. 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. In the simplest case, attention is assigned to the location corresponding to the peak of the attention map. However, this can result in an overly saccadic system. We therefore covertly moderate the attention map peaks 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 n_s times as salient as the next highest peak in the attention map; *clustered saliency* - attention is won by the view frame location about which n_c global peaks occur within p consecutive frames in a vicinity of radius r; *timeout* - if neither of the above winners emerge in t seconds, attention is given to the highest peak in the attention map since the previous fixation location was selected.

2.5 Processing Network

We adopt a client-server network processing architecture to allow concurrent serial and parallel processing. To minimise network bandwidth, to cope with the processing load of each frame, and to prevent repetition of computations, nodes in the structure are configured simultaneously as clients of processes preceding them in cue serialisation, and as servers to nodes following them. Trade-offs exist between splitting tasks into sub tasks, passing subtasks to additional nodes, and minimising network traffic. The best performing solution involves grouping serialised tasks on each server, and that as many operations are done on the image data on the same server as possible, so that there is minimal CPU idle time and minimal network traffic. The serial nature of cue computations means that there is often no additional gain possible in distributing tasks – in fact further network transfer of data between servers would slow performance and introduce additional latency. Fig.3 depicts the interconnectivity of the main processing nodes for comparison with a simplistic representation of the primate visual brain. At the lowest level, a dedicated video server obtains camera images for network distribution. A motion control server reports the head status and accepts remote motion control requests. A rectification server receives camera images and encoder status and distributes rectified images and rectification parameters to dependent nodes. Rectified U and V colour chrominance images are sent to colour centre-surround nodes for cue processing[11]. Intensity images are sent to the depth and flow processing servers. Foveal images are sent to the MRFZDF processing node. A spatial

[11] In primates, retinal ganglion colour opponent responses also propagate along separate pathways to intensity [1].

perception node receives depth and flow cues for augmentation into the occupancy grid and/or egosphere. The attentional nodes receive visual cues for production of the saliency, IOR and TSB maps, attention map and issues gaze control requests.

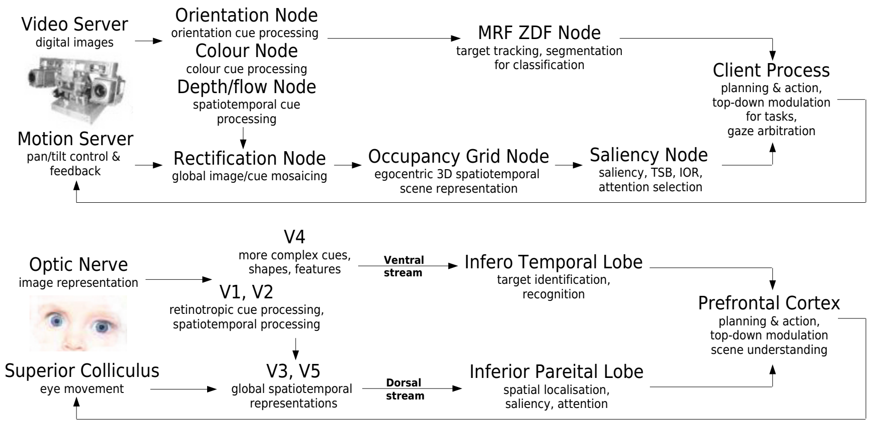
Video Server
digital images
Motion Server
pan/tilt control & feedback
Orientation Node
orientation cue processing
Colour Node
colour cue processing
Depth/flow Node
spatiotemporal cue processing
Rectification Node
global image/cue mosaicing
MRF ZDF Node
target tracking, segmentation for classification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

Fig. 3. 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). Feedback and minor pathways omitted.

2.6 System Behaviour

The robotic vision system preferentially directs its attention towards previously unattended salient objects/regions. Upon saccading to a new target, the MRF ZDF cue extracts the object (visual surface) that has won attention, maintaining stereo fixation on that object (smooth pursuit), regardless of its shape, colour or motion. Track is maintained until a more salient scene region is encountered, until IOR allows alternate locations to win fixation, or until variations in top-down attentional modulation yield an alternate peak location. During an attentional saccade, significant motion blur is observed, temporarily reducing image quality and affecting cue processing. In particular, the optical flow and disparity cues become excessively noisy. This noise can be misinterpreted in centre-surround processing as saliency. Excessive noise in optical flow calculations can also affect the propagation of dynamic IOR. To overcome this problem, the gaze moderation process broadcasts to the relevant processing nodes that a saccade is about to occur. Then, during the saccade, processing nodes can take appropriate action. For example, the propagation of IOR according to flow does not occur. Further, the MRF ZDF thread suspends issuing tracking commands until the saccade is complete[12].

During observation of a static scene, the interaction of IOR, saliency, and moderation typically induced a cyclical attentional scanpath where attention rotates through several of the most salient locations. In somewhat related work, Horowitz and Wolfe proposed that visual search is memoryless [11] - when elements of a search array were randomly re-organised while subjects searched for a specific target, search efficiency was not degraded. Performance gains for searches on a stable array would indicate memory use. However, this may just preclude perfect memorisation and does not necessarily preclude the possibility that the last few attended locations are remembered, in accordance with the limited lifespan of IOR. Other psycho-physical experimentation with static stimulus [12] suggested that a short-term attentional memory maintains information about salient visual features and their locations ("object files") across saccades, and that up to three or four object files may be retained. When several salient objects were present in front of the robotic system, they were re-attended at an approximately even rate, and in a largely cyclical, yet somewhat chaotic, order. This re-attention behaviour elicited by the robotic system is consistent with both of the stipulations above.

3 Psychophysical Trials

Similarities between the underlying robotic 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 (top, Fig. 4). Stimuli that may elicit emotional or significant cognitive responses were avoided. Participants were given a basic visual task (to count how many individual apples they saw amongst various fruit) while a non-intrusive gaze tracker (FaceLAB; bottom, Fig.4) recorded the path of their gaze (left, Fig.5). Each individual participated in only one trial. The scene could not be reproduced identically for each participant, but variables such as the rotation of the objects around string axes, and swinging, were similar in character across all trials. Although no two participants' gaze was expected to follow the same scanpath, we do expect to find some statistical similarities in terms of inter-individual gaze behaviours. Identical trials were conducted with the robotic system for statistical comparison to the human trial data.

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

[12] Interestingly, there exists a similar warning mechanism in biology: in neural recording studies with monkeys, scientists found that they could predict the occurrence of saccades by monitoring the activity of certain neurons [27, 6].

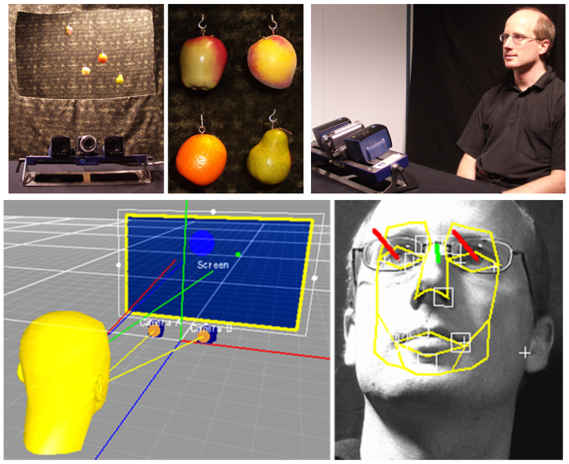

Fig. 4. Psychophysical trials (top): participant's view (left); trial stimuli (centre); non-intrusive gaze tracking (right). FaceLAB gaze tracking (bottom): extraction of gaze coordinates (left); and, modeling of head pose and gaze (right).

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. The main empirical observations during the human trials include [5]:

1. Gaze consistently saccades to the translating object.
2. During T, participants preferentially smoothly pursued translating stimulus.
3. Histograms of gaze velocities were strongly bimodal.
4. Saccade frequency was observed to decline during T, and increase during NT.
5. Saccade characteristics (such as velocity, distance) were *not* observed to vary significantly between T and NT.
6. Smooth pursuit characteristics *were* observed to vary significantly between T and NT.
7. Histograms of smooth pursuit distances show that a lower proportion of short smooth pursuit distances exist during T than NT.
8. The distribution of smooth pursuit gaze points during T correspond well to the paths of translating objects.

9. During NT, gaze frequented the locations corresponding to objects more than the background.
10. Re-attention periods were largely constant for all objects within an individual trial.
11. Velocity, position and smooth pursuit duration histograms exhibited inter-individual consistency.

Based upon these observations, 13 trial parameters (a non-limiting set) were extracted from each trial data log (left, Table 1): average smooth pursuit durations, distances, and velocities (for both T and NT - 6 parameters); average saccade distances and velocities (for both T and NT - 4 parameters); a saccade frequency parameter (for both T and NT - 2 parameters); and an average object re-attention period parameter (P) evaluated over all objects in a trial during NT (e.g, $P = 2.0$ represents that each object in the scene was re-attended on average once every 2.0 seconds, evaluated during NT periods where no objects are translating). To reduce the impact of participant mood/alertness, ratio parameters between T and NT were extracted from each trial providing 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).

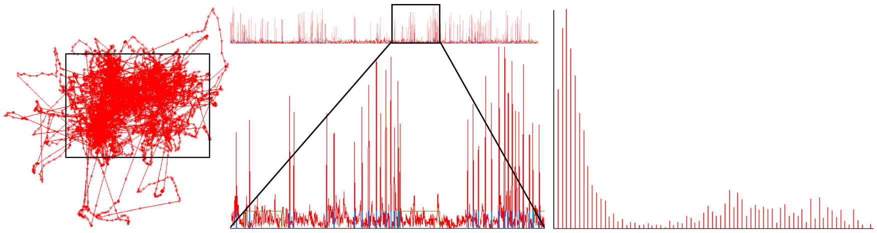

Fig. 5. Data for a single human trial (units ommited): 2D projection of complete gaze path with location of scene window (left); gaze velocity magnitude time-line (centre, above) with enlargement (centre, below) showing saccades (blue) and periods of object translation (green); and, histogram of velocity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

The seven parameters form the basis of the inter-individual statistical analysis. 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 [20] fail for most rate parameters unless less restrictive thresholds are chosen than recommended. Consequently, we *bootstrap*[13] [7] 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.

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	$P_{sd} = \mathrm{STD}(P_o)$

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

The bootstrapped inter-individual behavioural parameters demonstrate the following characteristic trends:

1. Smooth pursuit duration rate (Spt_r) varied significantly across participants, as characterised by the comparatively large bootstrapped standard deviation. This parameter is therefore largely dependent on the participant. There was a slight tendency for the parameter to increase during NT (suggesting a slight tendency for extended pursuit of translating stimuli) but the bootstrapped mean was centred at approximately 1.0.
2. Smooth pursuit distance rate parameter (Spl_r) and smooth pursuit velocity rate parameter (Spv_r) both consistently tended to decrease (< 1.0) during NT, commensurate with the tendency for participants to track translating stimuli. Additionally, the comparatively small bootstrapped standard deviations on these parameters characterise a generally similar decrease across all participants, and suggest that these parameters are largely scene-dependent.
3. Saccade distance rate parameter (Scl_r) consistently increased (> 1.0) during NT. The bootstrapped standard deviation in the parameter was comparatively large. This suggests some general scene dependency, but the increase depends largely on the participant.
4. Saccade velocity rate parameter (Scv_r) was approximately 1.0, suggesting that this parameter is not significantly dependent on the scene. The low/medial bootstrapped standard deviation in the parameter across participants is likely to reflect mechanical constraints (e.g, oculomuscular agility).

[13] ”Bootstraping” uses permutations of an available sample to generate many other samples with the same underlying PDF.

5. Saccade frequency rate parameter (Scf_r) consistently increased (> 1.0), characterising the tendency for the saccade rate to increase during NT across all participants. Moderate variance in this parameter across participants is shown statistically by the medial/large range in the parameter's bootstrapped standard deviation, suggesting the amount of increase is somewhat dependent on the participant.
6. The average object re-attention period during NT for each participant (P) varied significantly (STD(P)=1.92, calculated across all objects in all trials). However, object re-attention periods *for each participant* were significantly more constant (bootstrapped mean P_{sd} range of 0.12-0.52, significantly less than 1.92) as reflected in the small bootstrapped standard deviation.

3.2 Robotic Trials

Robotic trials were then conducted using the same trial apparatus and stimuli as for the human participants. The search task was effected by firstly recording colour chrominance samples from images of the target apple. These chrominance levels were used to set the desired search colours in the colour processing server node, whose output was weighted heavily in the construction of the saliency map. Additionally, the response of multiple orientations in the orientation processing server node were positively biased. In this manner, the attention system is predisposed to respond most strongly to small, round objects of a similar colour to the search target.

Before the first trial was conducted, system configuration settings (such as saliency map cue weights) were set by hand to mid-range values. After the first trial, configuration settings were iteratively adjusted such that the system was deemed likely to elicit behaviours more similar to human performance. For example, the first trial was noticeably more saccadic than the human trials. Predictions based on the system model were used to adjust the configuration settings to reduce the saccade rate - increasing the rate of accumulation of IOR over the fixation point, reducing the IOR decay rate of the entire dynamic IOR mosaic, and adopting more strict covert fixation moderation settings were predicted to lower the saccade rate. As per the human trials, distance-weighted velocity histograms of gaze path data were significantly bimodal. Ratio parameters, and the re-attention consistency parameter, were extracted from each robotic trial for comparison with the human rate parameter behavioural statistics.

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

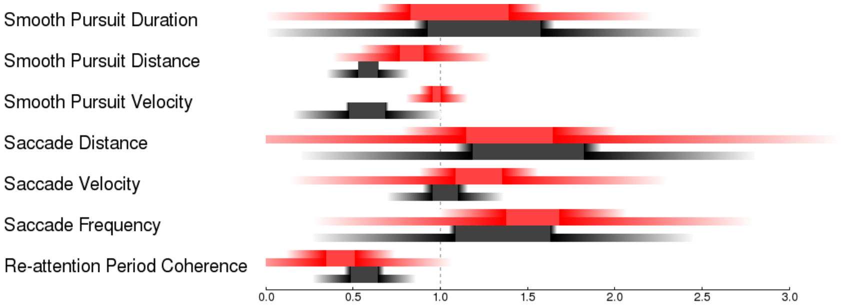

Fig. 6. Bootstrapped human (red) and robotic (black) inter-individual rate parameters. Distributions represent the rate change from periods where an object is translating (T) to periods where no objects are translating (NT). Each solid central bar region represents the bootstrapped 95% CI for the distribution 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.

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 [20] 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.

4.1 Parameter Sensitivity & Behavioural Variance

The bootstrapped robotic statistics may validly be used as metrics to assess the sensitivity (variances) in output behaviour to variations in input configuration settings. In the human trials, the largest variation in bootstrapped rate parameter distributions occurred in the smooth pursuit duration rate Spt_r (upper bound on 95% standard deviation CI 0.43), saccade distance rate Scl_r (0.85), saccade velocity rate Scv_r (0.49), and saccade frequency rate Scf_r (0.57) - suggesting that though the general trends in these parameters were the same across participants, the magnitude of change depends largely on the participant. Other human rate parameter distributions, including the smooth pursuit distance and velocities, exhibited lower variance - suggesting (as expected) that they may be more dependent on the repeatability of the scene than the participant. In the robotic trials, the largest variation in extracted rate parameter ranges also occurred in the saccade distance rate Scl_r (0.49), the smooth pursuit time Spt_r (0.46), and saccade frequency rate Scf_r (0.41). Saccade velocity rate Scv_r did not exhibit a variance as large as measured in the human trials, but again this is likely to be partially due to higher accuracy in velocity measurements using the robotic system's encoders (rather than the FaceLAB gaze estimation in the human trials), and the maximum velocity of the apparatus. Other than this instance, parameter variances were similar for both the robotic and human trials. Rate parameters that exhibited greatest variance in the robotic trials (Scl_r, Spt_r, Scf_r) suggest that these are more sensitive to configuration setting changes, rather than scene dependency. More robotic trials with stronger independence (randomly selected settings) would be required to confirm this hypothesis more conclusively. For both the robotic and human trials, the object re-attention period varied *across* trials, but coherence was demonstrated in the object re-attention period *within* each individual trial. The object re-attention period coherence parameter (P_{sd}) was *not* significantly sensitive to parameter variations. The average object re-attention period (P) within each trial *was* sensitive to configuration variations, as expected. Object re-attention periods were slightly less coherent across objects in the robotic trials (average standard deviation P_{sd} of 0.56) than the human trials (0.43). Nonetheless, standard deviations remained consistently low in both cases, and significantly lower than the standard deviation of the object re-attention period across all objects in all trials (1.19 for the robotic trials, 1.92 for human). The similar trends and trial parameter variances of the robotic and human systems further suggests behavioural consistency.

5 Conclusion

The trials were not tailored to determine the correct object re-attention period, IOR radius, IOR decay rate, tracking periods, or configuration settings. These parameters are likely to differ greatly across human participants, and even over time for a particular individual. Even though the 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 comment on emergent gaze behaviours observed in the robotic trials for comparison with benchmarks obtained from the human trials. 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, and accordingly similar sensitivity to parameter variations, suggests similar performance does not rely purely upon the selection of configuration settings. Rather, 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 when subjected to the same visual stimuli.

References

1. M. Dacey, “Circuitry for color coding in the primate retina,” in *Proc. Nat. Acad. Sci.*, 1996, pp. 93:582–588.
2. A. Dankers, N. Barnes, and A. Zelinsky, “Active vision - rectification and depth mapping,” in *Aust. Conf. Robotics and Automation*, 2004.
3. ——, “Active vision for road scene awareness,” in *IEEE Intelligent Vehicles Symposium*, 2005.
4. ——, “Mapzdf segmentation and tracking using active stereo vision: Hand tracking case study,” in *Comp. Vis and Im. Understanding*, 2006.
5. A. A. Dankers, “Real-time synthetic primate vision,” in *PhD Thesis*, 2007.
6. M. C. Dorris, M. Pare, and D. P. Munoz, “Neuronal activity in monkey superior colliculus related to the initiation of saccadic eye movements,” in *J. Neuroscience*, 1997.
7. B. Efron and R. J. Tibshirani, “An introduction to the bootstrap,” in *Chapman and Hall*, 1993.
8. C. Gilbert, M. ito, M. Kapadia, and G. Westheimer, “Interactions between attention, context and learning in primary visual cortex,” in *Vision Res.*, 2000, pp. 40:1217–1226.
9. 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.

10. R. Hartley and A. Zisserman, *Multiple View Geometry in Computer Vision, Second Edition.* Cambridge University Press, 2004.
11. T. S. Horowitz and J. M. Wolfe, "Visual search has no memory," in *Nature*, 1998.
12. D. E. Irwin and G. J. Zelinsky, "Eye movements and scene perception: Memory for things observed," in *Perception and Psychophysics*, 2002.
13. L. Itti, "Models of bottom-up attention and saliency," in *Neurobiology of Attention*, 2005.
14. M. H. Johnson, "Developmental cognitive neuroscience, vol. 1, 3 ed. malden," in *MA and Oxford UK: Balckwell Publisher Inc.*, 1997.
15. S. Kagami, K. Okada, M. Inaba, and H. Inoue, "Realtime 3d depth flow generation and its application to track to walking human being," in *IEEE International Conf. on Robotics and Automation*, 2000, pp. 4:197–200.
16. E. R. Kandel, J. H. Schwartz, and T. M. Jessell, "Principles of neural science, 4th edition," in *McGraw-Hill Medical*, 2000.
17. R. M. Klein, "Inhibition of return," in *Trends in Cognitive Sciences*, 2000.
18. P. Kovesi, "Phase congruency detects corners and edges," in *Aust. Patt. Rec. Soc.*, 2003, pp. 309–318.
19. E. Merriam, C. Genovese, and C. Colby, "Spatial updating in human parietal cortex," in *Neuron*, 2003, pp. 39:351–373.
20. T. M. Mitchell, "Machine learning," in *McGraw-Hill*, 1997.
21. V. Navalpakkam, M. Arbib, and L. Itti, "Attention and scene understanding," in *Neurobiology of Attention*, 2005.
22. P. Neri, H. Bridge, and D. J. Heege, "Stereoscopic processing of absolute and relative disparity in human visual cortex," in *Neurophysiol. 92: 18801891*, 2004.
23. S. Nieuwenhuis and N. Yeung, "Neural mechanisms of attention and control: losing our inhibitions?" in *Nature*, 2005, pp. 8:1631–1633.
24. 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.
25. J. H. Reynolds, T. Pasternak, and R. Desimone, "Attention increases sensitivity of v4 neurons," in *Neuron, 26 3:703-714*, 2000.
26. 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.
27. L. Sugrue, G. S. Corrado, and W. T. Newsome, "Choosing the greater of two goods: Neural currencies for valuation and decision making," in *Neuroscience*, 2005.
28. Sun and Bonds, "Two-dimensional receptive field organization in striate cortical neurons of the cat," in *Vis Neurosci.*, 1994, pp. 11: 703–720.
29. A. Trieisman and G. Gelade, "A feature-integration theory of attention," in *Cogn. Psychol.*, 1980, pp. 12:97–136.
30. H. Troung, "Active vision head," in *Thesis, Australian Nat Univ.*, 1998.
31. W. E. Vinje and J. L. Gallant, "Sparse coding and decorrelation in primary visual cortex during natural vision," in *Science 287:12731276*, 2000.
32. 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.