A Reactive Vision System: Active-Dynamic Saliency

Andrew Dankers[1,2], Nick Barnes[1,2], and Alex Zelinsky[3]

[1] National ICT Australia[4], Locked Bag 8001, Canberra ACT Australia 2601
[2] Australian National University, Acton ACT Australia 2601
{andrew.dankers, nick.barnes}@nicta.com.au
[3] CSIRO ICT Centre, Canberra ACT Australia 0200
alex.zelinsky@csiro.au

Abstract. We develop an architecture for reactive visual analysis of dynamic scenes. We specify a minimal set of system features based upon biological observations. We implement feature on a processing network based around an active stereo vision mechanism. Active rectification and mosaicing allows static stereo algorithms to operate on the active platform. Foveal zero disparity operations permit attended object extraction and ensures coordinated stereo fixation upon visual surfaces. Active-dynamic inhibition of return, and task dependent biasing result in a flexible, preemptive and retrospective system that responds to unique visual stimuli and is capable of top-down modulation of attention towards regions and cues relevant to tasks.

Key words: Realtime, Dynamic, Active, Stereo, Vision, Saliency

1 Introduction

We work towards a reactive synthetic active vision system based upon observations of biology. At the heart of the system is a vision mechanism that exhibits a mechanical performance similar to that of primates. CeDAR (left, Fig.1), the Cable-Drive Active-Vision Robot [15], incorporates a common tilt axis and two independent pan axes separated by a baseline of $30cm$. All axes exhibit a range of motion of greater than 90^o, speed of greater than $600^o s^{-1}$ and angular control resolution of 0.01^o. Accordingly, we develop a vision processing framework that takes much inspiration from primate vision. It is capable of detecting and responding to unique visual events, and of performing a variety of basic visual tasks.

We specify features considered necessary for the system (Sect.2). We describe the implementations of components of the system (Sect.3). We consider system structure, such that component dependencies are preserved when distributing tasks over a processing network (Sect.4). We present basic experiments and provide sample system output (Sect.5), before concluding (Sect.6).

[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.

Fig. 1. Left: CeDAR, active vision apparatus. Right: Active rectification mosaic.

2 System Specification

It is desirable that system operation does not contradict known properties of the primate vision system. Where possible, we use observations of biology to specify methods to deal with active cameras (Sect.2.1), to define features of fixation control (Sect.2.2), and to choose a relevant set of early visual cues (Sect.2.3).

2.1 Egocentric Reference Frame

Rectification: Camera relations need to be determined for each image pair such that static stereo algorithms can be used with active cameras. Rectification incorporates removal of lens effects such as barrel distortion.

Spatial Continuity: Monkeys retain a short term memory of attended locations across saccades by transferring activity among spatially-tuned neurons within the intraparietal sulcus [10], thus retaining accurate retinotropic representations of visual space across eye movements. Accordingly, the synthetic system should exhibit a pseudo-retinotropic coordinate system so that the relations between left and right and between successive images are known, despite camera motions such as smooth pursuit and saccade.

2.2 Fixation

Coordination: For humans, one cue used to extract the boundary of an attended object is *zero disparity*. An attended object appears at near identical positions in left and right retinas, whereas the rest of the scene usually does not. That is, the attended object is at zero disparity. During stereo fixation, the foveas are aligned in a truly coordinated manner.

Segmentation: Long range excitatory connections in V1 appear to enhance responses of orientation selective neurons when stimuli extend to form a contour [6]. Monkeys exhibit vigorous responses elicited by small laboratory stimuli

in isolation, compared to sparse neuronal activity when viewing broad scenes. Accordingly, the synthetic system should respond to the contours of the object upon which fixation occurs. This is a foveal response, coupled to the coordinated fixation feature.

Dynamic Inhibition of Return (IOR): Monkeys maintain accurate representations of visual space across eye movements. A short-term effect prevents previously attended stimuli from being immediately re-attended. In this sense, it is a retrospective response that depends upon past observations.

Task Biasing: The prefrontal cortex implements attentional control by amplifying task-relevant information relative to distracting stimuli [11]. Bias may be preempted for regions of the scene not currently in view, but whose position relative to the current fixation point is known in retinotropic coordinates.

2.3 Saliency

Pre-attentive feature computation occurs continually in primates across the entire visual field: a neuron will fire vigorously even if the animal is attending away from that neuron's receptive field, or if the animal is anesthetized [14]. Cue contrast is paramount in saliency, not local absolute cue levels [12]. Higher level cues also contribute to saliency: eye trackers have been used to observe that humans preferentially fixate upon regions with multiple orientations [18]; another example is the neural critical collision response observed in pigeons [17]. Neurons at early stages in the primate visual brain are tuned to simple features like intensity contrast, colour opponency, orientation, motion, and stereo disparity. Desirable synthetic system cues include: depth, optical flow and depth flow, colour, intensity, orientation, and collision criticality.

3 Implementation of Specified Components

3.1 Fixation and Egocentric Reference Frame

Active Rectification: In [4] , we described a method to rectify camera barrel distortions and to actively enforce *parallel epipolar geometry* [7]. This work enables online epipolar rectification of the image sequences and the calculation of the shift in pixels between consecutive frames from each camera, and between the current frames from the left and right cameras. We construct globally epipolar rectified mosaics of a scene as the cameras move, in realtime (right, Fig.1).

Zero Disparity: In light of primate vision, we define a static synthetic fovea approximately the size of a fist held a distance of $60cm$ from the camera. For our cameras, this corresponds to a region of about 100x100 pixels. A robust *zero disparity filter* (ZDF) has been formulated [5] to optimally identify objects that map to image frame pixels at the same coordinates in the left and right camera foveas, regardless of foreground or background clutter. Fig.2 shows sample output of the MRF ZDF cue.

Fig. 2. a) MRF ZDF output (right) with left and right input (respectively) and foveal processing regions. b) Robust performance – segmentation of attended hand from face in near background (top left); a distracting hand (bottom left); and an occluding hand a distance of $3cm$ from the tracked hand $2m$ from the cameras (top right). If closer than $3cm$, they are jointly segmented (bottom right).

3.2 Saliency Cues

Cue synthesis is subject to real-time performance constraints, so cues are implemented with processor economy in mind. YUV[4] channels are processed. Some serialisation in cue processing is required to meet dependencies (Fig.5, a).

Intensity Uniqueness: Neurons tuned to intensity centre-surround produce a response that can be synthesized using a *difference-of-Gaussian* (DoG) approximation [8]. In a manner similar to [16], we create a Gaussian pyramid from the intensity image. Successive images in the pyramid are down-sampled by a factor of two (n times), and each is convolved with the same Gaussian kernel. To obtain DoG images, the Gaussian pyramid images are upsampled (with bi-linear interpolation) to the original size and then combined. Combination involves subtracting pyramids at coarser scales C_n from those at finer scale C_{n-c}. We consider two levels of interaction, immediate neighbours $C_n - C_{n-1}$, and second neighbours $C_n - C_{n-2}$, to obtain a DoG pyramid with $n-3$ entries. Finally, the $n-3$ entries are added to obtain a map where the most spatially unique region emerges with the strongest response. The borders of the image equate to an edge that would otherwise produce a significant step response in uniqueness computations. Prior to processing, a smooth edge transition is enforced using a windowing function.

Colour Uniqueness: U and V chrominance uniqueness are computed as per

[4] YUV: Y represents the intensity channel, U and V are colour chrominance channels.

intensity, then combined by addition.

Optic Flow: The translation from the current to previous frame is known in mosaic coordinates. Optic flow is calculated on the overlapping region of consecutive view frames in the mosaic. This allows estimation of horizontal and vertical scene flow in the mosaic reference frame that is independent of the motion of the cameras. A *sum of absolute differences* (SAD) flow operation [1] is used. We obtain four maps: horizontal and vertical flows in each camera. Centre-surround uniqueness is determined for all four maps. We down-sample images before computing flow for processor economy.

Depth and Depth Flow: The epipolar rectified mosaics allow us to search for pixel disparities along horizontal scan-lines only. We search the neighboring ± 16 pixels in the second image for a correspondence to the candidate pixel location in the first image. We conduct the disparity search in the overlapping region of current left and right frames only. The velocities of visual surfaces in the depth direction are calculated using an approach similar to that of [9]. Centre-surround uniqueness is determined for depth and depth flow.

Orientation Uniqueness: Strong local interactions between separate orientation filters have been characterised via neuronal correlates [8]. A winner-take-all competition is activated amongst neurons tuned to different orientations and spatial frequencies within one cortical hypercolumn [2]. A synthetic response is achieved using complex log-Gabor convolutions over multiple scales within each orientation. The log-Gabor response is similar to the impulse response observed in the orientation sensitive neurons in cats [13]. We obtain orientation response maps for each orientation and scale. Within each orientation, we sum all scale responses.

Critical Collision Cue: Visual surfaces on an instantaneous trajectory leading towards the visual apparatus elicit the strongest response. The cue is defined:

$$\frac{||p||}{||v||}(1 - (-nv \cdot np)), \tag{1}$$

where $p = (x, y, depth)$ and $v = (flowx, flowy, depthflow)$ are position and velocity vectors, the dot represents the dot product, and $nv = v/|v|$, and $np = p/|p|$ are unit vectors.

3.3 Dynamic Fixation

We introduce three intermediary maps such that a fixation map can be determined for the active cameras with dynamic scenes. The three maps include a saliency map, an *IOR* mosaic, and a *task-dependent spatial bias (TSB)* mosaic.

Saliency: Center-surround cues are weighted and added into a single saliency

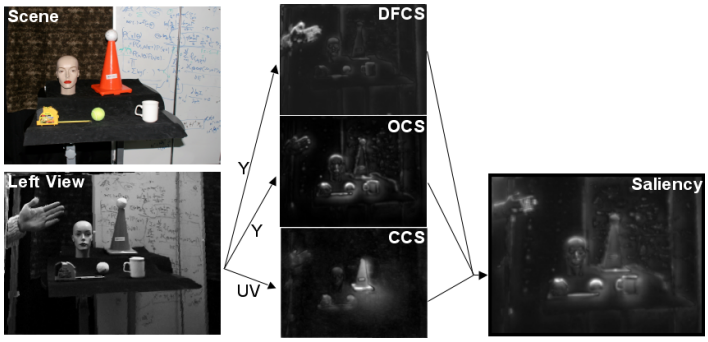

Fig. 3. Server cue uniqueness responses contribute to saliency.

map for each camera, and the result is normalised (Fig.3).

IOR: A Gaussian kernel is added to the region around the current fixation point in an IOR accumulation mosaic, every frame (Fig.4, b). Expanding upon this for dynamic scenes, accumulated IOR is propagated according to optic 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. We decrement IOR over time according to decay rate I_d, so that previously inhibited locations eventually become uninhibited. Faster I_d decay means more frequent saccades. This rate can be modulated by higher level client processes. IOR is a retrospective response that depends upon previous observations. For a given head pose, the mosaic reference frame remains static with respect to the world, and as such, regions of the mosaic not in the current view frame may remain inhibited until decayed or next attended. Fig.4 (a) shows the interaction between dynamic IOR and saliency.

TSB: The prefrontal cortex implements attentional control by amplifying task-relevant information relative to distracting stimuli [11]. We introduce aTSB mosaic (Fig.4, b) that can be dynamically tailored according to tasks. For example, when driving a car, we tend to keep our eyes on the road, and as such we bias the lower half of the mosaic where we would expect to find the road. TSB can be preempted for regions not in the current view frame.

Target Selection and Pursuit: In its simplest form, attention is assigned to the location corresponding to the peak of the fixation map. However, this can result in an overly saccadic system. We therefore moderate the winning locations before the winner of fixation is selected: *Supersaliency*: a view frame coordinate immediately wins attention if it is n_s times as salient as the next highest peak. *Clustered Saliency*: attention is won by the view frame location about which n_c global peaks occur within p consecutive frames. *Timeout*: if neither of the above winners emerge in t seconds, attention is given to the highest peak in the fixa-

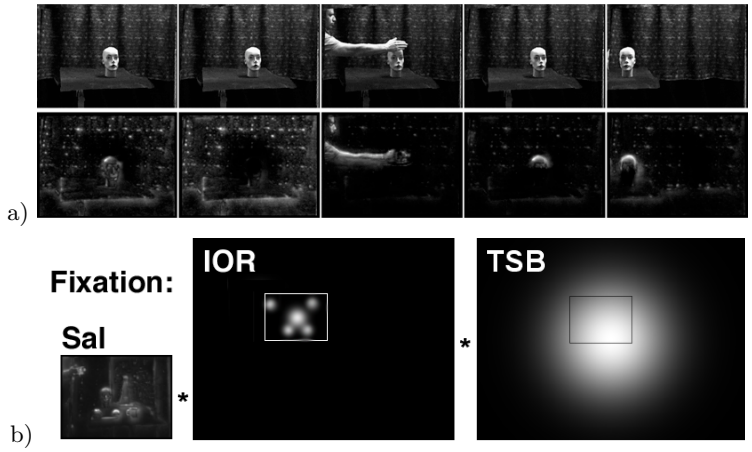

Fig. 4. a) Dynamic IOR (from left) 1. The head on trolly moves into fovea, initially uninhibited; 2. After time it becomes inhibited; 3. A salient hand enters fovea; 4. IOR on forehead is reset by occlusion; 5. Trolley and head move out of fovea, taking associated IOR pattern. b) Fixation is the product of view frame saliency, dynamic IOR, and TSB. This radial TSB could represent a forwards search task. The gradient across the view frame induced by the TSB enhances saliency towards the centre of the mosaic.

tion map since the last winner. Once attention is won, MRF ZDF segmentation ensures fixation upon the object or surface that won attention, even it if moves, until the next winner is selected.

4 System Structure

We adopt a client-server architecture to allow concurrent serial and parallel processing. At the lowest level, a rectification server distributes rectified images and rectification parameters to dependent nodes. Biological evidence suggests that colour is treated in separate regions in the brain to intensity [3]. U and V colour chrominance images for both the left and right images are sent to the colour centre-surround server (CCS) for 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 (Fig.5, a), and as servers to nodes following them. Each node is a dual CPU hyper-threaded 3GHz PC with four virtual processors. 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 gain possible in distributing the task – in fact further network transfer of data between servers would slow per-

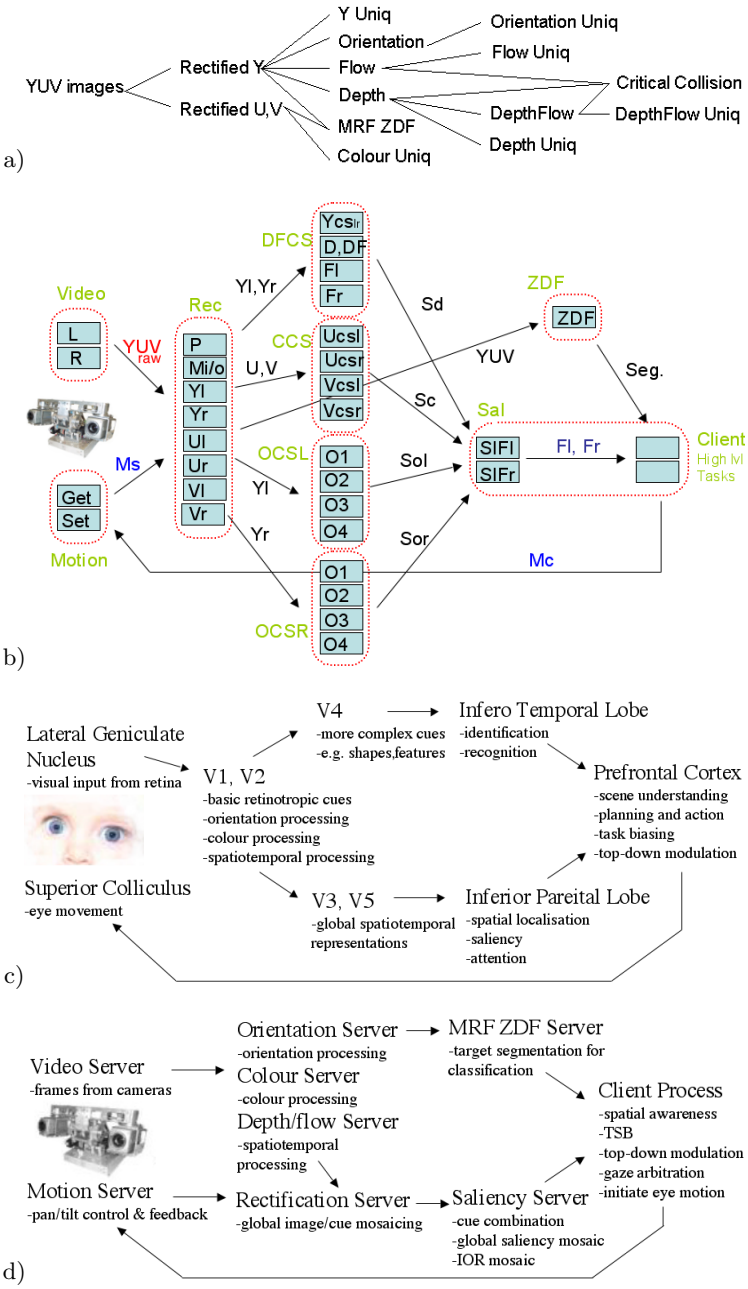

Fig. 5. a) Synthetic cue dependencies. b) System block diagram - Dotted lines surround physical PCs. Boxes show processing threads. Arrows show major data flows: motion status (Ms), motion commands (Mc), saliency maps (Sd,Sc,Sol,Sor), fixation maps (Fl,Fr), target segmentation (Seg.) and image channels (Y,U,V). c) Broad interactions in primate visual brain. d) The synthetic vision system. Modulation feedback pathways omitted.

formance. Fig.5 (b) is a block diagram summarising data flow occurring between each client/server node in the processing network. Fig.5 (c, d) shows a broad model of the major interactions in the primate visual brain, and the synthetic vision system. It is noted that the synthetic structure bears a good resemblance to primates.

Fig. 6. Sample behavior: (from left) 1. Attention shifts to head from inhibited base of cone, forehead is segmented from background in fovea; 2. Attention returns from inhibited head to top of cone, cone is segmented; 3. Attention shifts from inhibited cone to mug, mug is segmented.

5 Experiments

The synthetic 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 that has won attention, maintaining stereo fixation on that object (smooth pursuit), regardless of its shape, colour or motion. Attention is maintained until a more salient scene region is encountered, or until IOR allows alternate locations to win fixation (Fig.6). *Demonstration footage* available at:

`http://rsise.anu.edu.au/~andrew/icvs07`

6 Conclusion

By specifying system properties similar to those observed in nature, we have developed a synthetic active visual system capable of detecting and reacting to unique and dynamic visual stimuli, and of being tailored to perform basic visual tasks. Foveal zero disparity operations permit attended object extraction and ensures coordinated stereo fixation upon visual surfaces. Active-Dynamic IOR means that a short term memory of previously attended locations can be

retained to influence attention retrospectively. Spatial and cue biassing based on observations and prior knowledge allow preemptive top-down modulation of attention towards regions and cues relevant to tasks. These features result in a reactive vision system and the emergence of intelligent attentional behaviors.

References

1. J. Banks and P. Corke, "Quantitative evaluation of matching methods and validity measures for stereo vision," *IEEE International Journal of Robotics Research*, vol. 20, no. 7, 1991.
2. M. Carrasco, C. Penpeci-Talgar, and M. Eckstein, "Spatial convert attention increases contrast sensitivity across the csf: support for signal enhancement," in *Vision Res.*, 2000, pp. 40:1203–1215.
3. M. Dacey, "Circuitry for color coding in the primate retina," in *Proc. Nat. Acad. Sci.*, 1996, pp. 93:582–588.
4. A. Dankers, N. Barnes, and A. Zelinsky, "Active vision - rectification and depth mapping," in *Australian Conf. on Robotics and Automation*, 2004. [Online]. Available: http://www.araa.asn.au/acra/acra2004/
5. ——, "Mapzdf segmentation and tracking using active stereo vision: Hand tracking case study," in *Comp. Vis and Im. Understanding*, 2006.
6. 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.
7. R. Hartley and A. Zisserman, *Multiple View Geometry in Computer Vision, Second Edition*. Cambridge University Press, 2004.
8. L. Itti and C. Koch, "Feature combination strategies for saliency-based visual attention systems," in *J. Electronic Imaging*, 2000.
9. 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.
10. E. Merriam, C. Genovese, and C. Colby, "Spatial updating in human parietal cortex," in *Neuron*, 2003, pp. 39:351–373.
11. S. Nieuwenhuis and N. Yeung, "Neural mechanisms of attention and control: losing our inhibitions?" in *Nature*, 2005, pp. 8:1631–1633.
12. H. Nothdurf, "Texture discrimination by cells in the cat lateral geniculate nucleus," in *Exp. Brain Res*, 1990, pp. 82:48–66.
13. Sun and Bonds, "Two-dimensional receptive field organization in striate cortical neurons of the cat," in *Vis Neurosci.*, 1994, pp. 11: 703–720.
14. S. Treue and J. Maunsell, "Attentional modulation of visual motion processing in cortical areas mt and mst," in *Nature*, 1996, pp. 382:539–541.
15. H. Truong, S. Abdallah, S. Rougeaux, and A. Zelinsky, "A novel mechanism for stereo active vision," in *Australian Conf. on Robotics and Automation*, 2000.
16. A. Ude, V. Wyart, M. Lin, and G. Cheng, "Distributed visual attention on a humanoid robot," in *Report*, 2005.
17. D. Wylie, W. Bischof, and B. Frost, "Common reference frame for neural coding of translational and rotational optic flow," in *Nature*, 1998, pp. 392:278–282.
18. C. Zetzsche, "Investigation of a sensorimotor system for saccadic scene analysis: an integrated approach," in *5th Intl. Conf. Sim. Adaptive Behav.*, 1998, pp. 5:120–126.