1 Computational vision in neural and machine systems

Michael Jenkin and Laurence Harris

1.1 Introduction

The ability to process visual information streams is a critical requirement for both biological systems as well as for a wide variety of robotic devices. The fundamental need for effective visual information processing in biological systems is illustrated in Figure 1.2. These three snapshots show a cheetah emerging from the long grass in Tanzania. Being able to discern the shape emerging from the tall grass *early* is a critical survival skill. Biological systems that are unable to process the visual information in a timely fashion are unlikely to succeed in the wild. Similarly, in the machine vision domain, timely effective processing of visual information is often key. Figure 1.3 shows the AQUA robot, a visually guided amphibious robot that is capable of unsupervised operation (Dudek *et al.*, 2005). This vehicle relies primarily on visual information obtained from forward facing cameras to reason about its external environment. Without the ability to process its multiple video camera inputs in a timely fashion, the robot would be unable to operate.

In 1991 (over 15 years ago), the Centre for Vision Research at York University held the first in what has become a bi-annual conference on vision. The 1991 conference – which resulted in the book *Spatial Vision in Humans and Robots* – examined how biological and machine systems address the task of processing the rich visual field in order to recover information about the spatial surround. Fifteen years later, the York Vision Conference has re-visited this fundamental issue: the relationship between computational models of visual information processing and research into biological visual information processing.

The past fifteen years have seen astounding advances in both artificial and biological vision. Driven (at least in part) by advances in the technology available to explore how biological systems process visual information, and the orders of magnitude performance improvements in the computational power that can be brought to the task of

Computational Vision in Neural and Machine Systems, ed. L. Harris and M. Jenkin. Published by Cambridge University Press.

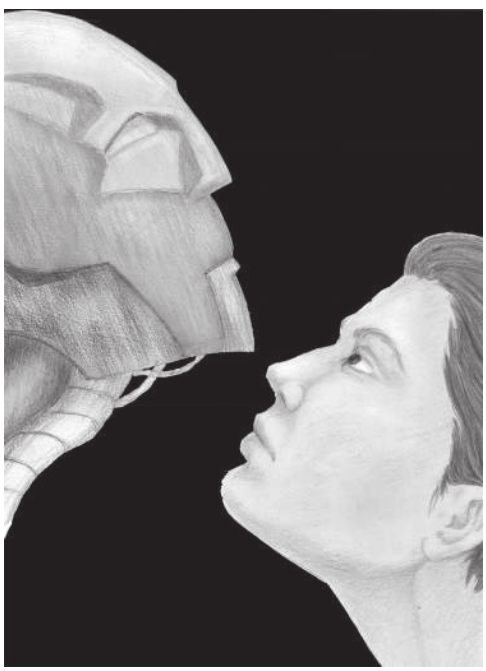

Figure 1.1. Computational vision in neural and machine systems. Appears with the kind permission of Emma Jenkin.

processing visual imagery, visual models have advanced from processing a single image viewed in isolation to a stream of embedded visual information processing within an ongoing spatio-temporal relationship. In the computational field, this has lead to the consideration of visual information processing not as the evaluation of an isolated static image, but rather as the task of processing an image stream within the context of some wider task. In the biological fields, this has lead to a wide range of advances including the emergence of models of multi-modal fusion of information from different perceptual systems.

As visual information processing is considered within a temporal context, many of the problems that occur "naturally" in the biological community become apparent in the computational one. This includes tasks such as integrating information from multiple views, searching for specific objects within a wide visual display, and attending to salient features within the environment.

David Marr (1982) distinguished three levels of visual processing: computational, algorithmic, and implementational. His computational level consisted of a description of what computation needs to be performed and what information is available to perform the computations on. His algorithmic level specified how the computational level might be performed. Algorithms performed biologically are likely to be very different from those performed on a computer. For example there are many levels of parallel

Figure 1.2. Vision is a critical perceptual ability for many biological systems. Early detection of visual events – such as that depicted above – can be essential for an individual's survival.

processing in the brain, which is unusual in a digital computer. The implementation level is the act of performing the selected algorithm, either in the brain or in a digital computer. This book places more emphasis on the first of Marr's three stages, outlining the principles of the computational processes to be performed with less emphasis on the actual algorithms that might be employed to run them.

This volume is divided into three parts, centred around the topics of dynamical systems; attention, motion and eye movements; and stereo vision. Dynamical systems deals with adaptation, motion detection, robotic vision systems, shape recovery from image sequences, and the reconstruction of objects from parts and attributes processed

separately. The section on attention deals with attention and action, visual search in clutter, the memory of visual features accross saccades, and modelling gaze in natural images. Finally, the section on stereo describes a number of algorithms and approaches that reflect the current state of the art in stereo vision algoirthms and models of stereo information processing.

In each of these sections we find papers that examine spatial information processing and how it interacts with the temporal domain. In Part I, for example, Norma Graham and Sabina Wolfson examine specific adaptation processes in human visual information processing. They question how various levels of visual information processing adapt to the absolute levels of illumination that are available and the time course of this adaptation. In Part II, Steven Prime, Matthias Niemeier, and Douglas Crawford examine how visual information is maintained across visual saccades, a problem that is critical to biological systems that utilize eye movements to integrate larger portions of the visual field than are available in a single gaze, and which is also critical in machine systems which must use camera and vehicle motion to deal with the limited field of view of existing camera technologies. Finally in Part III, Jane Mulligan examines how stereo image processing can be made sufficiently "computationally efficient" that it can be embedded within machine vision systems and used as a building block for telepresence systems.

Beyond these three examples we find a collection of chapters that seek to address spatial vision information processing in both the computational and biological fields. These chapters illustrate just how detailed our understanding of basic visual information processing has come, and how much remains to be discovered. They also demonstrate how similar the problems are that are encountered by biological and computational systems and how similar are the underlying information processing models (algorithms). Much has been accomplished in the fifteen years since the first York Vision Conference on Spatial Vision in Humans and Robots. As we observed in the introduction to the book that arose from that conference, the two communities can learn a great deal from each other. That observation seems just as true today.

1.2 The CD-ROM

Enclosed with this volume is a CD-ROM that contains video, colour imagery, and other digital media associated with the text. A complete copy of this volume in PDF format can also be found on the CD-ROM. The material on the CD-ROM can be accessed using a standard browser (such as Internet Explorer or Firefox). Videos on the CD-ROM are viewable with Quicktime, while viewing of the presentations on the CD-ROM will require a PowerPoint viewer.

References

Dudek, G., Jenkin, M., Prahacs, C. *et al.*, (2005). A visually guided swimming robot. *Proc. IROS 2005*. Edmonton, Alberta.

Harris, L. and Jenkin, M. (1993). *Spatial Vision in Humans and Robots*. New York:

Figure 1.3. The AQUA Robot. A visually guided amphibious robot (Dudek *et al.*, 2005).

Cambridge University Press.

Marr, D. (1982) *Vision: a Computational Investigation into the Human Representation and Processing of Visual Information.* San Francisco: W. H. Freeman.

Part I

Dynamical systems

2 Exploring contrast-controlled adaptation processes in human vision (with help from *Buffy the Vampire Slayer*)

Norma Graham and S. Sabina Wolfson

We have been interested for many years in intermediate levels of visual processing: levels which are lower than the perception of "objects" and "scenes" but higher than the pointwise processing of the retina and LGN.[1] Many of these intermediate processes are concerned with the initial analyses of pattern and form. Much of their action can be well modeled by what are technically linear operations – in particular, multiple analyzers sensitive to different ranges of spatial frequency and orientation.[2] However, some of their action cannot be modeled this way as it is fundamentally nonlinear. We have recently become interested in the dynamics of these intermediate nonlinear processes, and more particularly in questions about how the visual system sets its sensitivity based on the recent history of stimulation.

This chapter affords us an opportunity to be informal and to relate past work to present work in ways that are uncommon in journal papers. We are happy to take advantage of this opportunity. We will use informal speech and explanations and also personal anecdotes. And we will give many fewer references to published literature than is our wont, but instead will try to guide the reader to places where such references can be found.

Computational Vision in Neural and Machine Systems, ed. L. Harris and M. Jenkin. Published by Cambridge University Press.

[1] The terms "lower" and "higher" are only approximate, of course, since information travels "downstream" as well as "upstream." Lennie (1998) presents interesting hypotheses – and an overall view – about the function and nature of the processes that have physiological substrates from V1 up to V4 and MT.

[2] The psychophysical research on these multiple analyzers, and a small amount of the physiological research, is described in Graham (1989) and summarized in Graham (1992).

Two sets of psychophysical experiments – and the models that were tested by their results – are described in this chapter. The first is described briefly and the second at length.

The first set was designed to investigate behaviorally the dynamics of luminance-controlled processes like light adaptation in the retina or LGN. Strictly speaking, these processes are lower than the level that we have been most interested in and were done with a third major collaborator, Don Hood, who is very interested in that level. Further, this set is already published for the most part. Thus we will describe it quite briefly. However, we do describe it because it both inspired the second set and also gave us distinct expectations about how the second set would turn out.

The second set of experiments was designed to investigate the dynamics of contrast-controlled processes. We started out to study one such process that had proved necessary to explain our previous results with textured patterns (done in collaboration with other investigators, in particular Jacob Beck and Anne Sutter). But the results of this second set of experiments ended up suggesting the existence of an entirely different contrast-controlled process, and one that we had not previously even imagined. This second set of experiments and the new process they suggested will be the focus of most of this chapter.

2.1 Dynamics of luminance-controlled adaptation processes (light adaptation)

2.1.1 Flickering the luminance of spatially homogeneous backgrounds and measuring thresholds for superimposed luminance probes

The first set of experiments, that we will only discuss briefly here, was intended to investigate the dynamics of luminance-controlled adaptation processes (e.g. light adaptation in the retina). Figure 2.1 shows the spatial and temporal characteristics of this paradigm, which is often called the probed-sinewave paradigm.

In probed-sinewave experiments, the luminance of a spatially homogeneous background is flickered sinusoidally in time during each trial. At some point during the trial a luminance-defined probe is introduced (of intensity ΔI, an increment in the figure, but decrements have been used as well). It is typically a smaller disk in the middle of the flickering background.

You can see movies of these stimuli on the CD-ROM accompanying this book. Video 1 shows the flickering background disk by itself. Video 2 shows the flickering background disk with a probe increment introduced.

Results from one typical observer from one study are shown in Figure 2.2 (with separate frequencies of flickering background in separate panels) and then again in Figure 2.3 (with the results at different frequencies superimposed in one panel). Probe threshold is plotted as a function of phase, and, to help show trends in results, the phases are plotted through two cycles on the horizontal axis. The results in Figures 2.2 and 2.3 show typical features of experimental results from this paradigm. Of particular

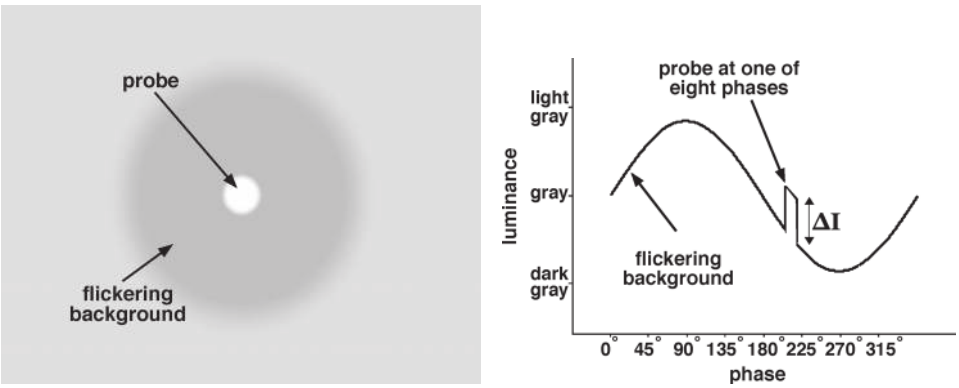

Figure 2.1. The probed-sinewave paradigm used to study luminance-controlled adaptation (light adaptation) processes. Spatial paradigm on left, and temporal paradigm on right. The luminance in the probe ΔI is adjusted until the observer can just discriminate between the background-alone and the background-plus-probe.

importance to the story here, note the large general increase in probe threshold magnitude (the upward displacement of the curves) as the background's flicker frequency is raised from lower (lighter and thinner lines and symbols) to higher (darker and bigger lines and symbols). The probe thresholds decrease at still higher frequencies – not shown here – until they are back down on average to the same level as at very low frequencies.

Many other studies – using different conditions and different observers – have been done by various groups of investigators. While the studies differed among themselves in a number of ways, they all showed a big general increase in probe threshold magnitude as the frequency of the flickering background increased from low to middling. (Many were compared in Graham, Wolfson, and Chowdhury, 2001.)

2.1.2 What do the results imply for models of light adaptation?

This empirical result – the general increase in probe threshold with increase in background flicker frequency – has turned out to be very powerful in discriminating among different models, or, more generally, in discriminating among different ideas of how light adaptation might work. Indeed Hood *et al.* (1997) showed that this empirical result completely rules out a large class of previously successful models containing the best features of the two earlier modeling traditions (the merged models of Graham and Hood, 1992). This empirical result could not, however, immediately rule out in the same dramatic way a new model suggested by Wilson (1997). The Wilson model, based on explicit physiological pieces, could be trivially modified to do a satisfactory job to at least a good first approximation (Hood and Graham, 1998, as was subsequently shown also with a fuller set of results by Wolfson and Graham 2000, 2001a). But the Wilson model has some drawbacks (see discussion in Wolfson and Graham, 2001a,b). The best current candidate in our opinion is the more abstract

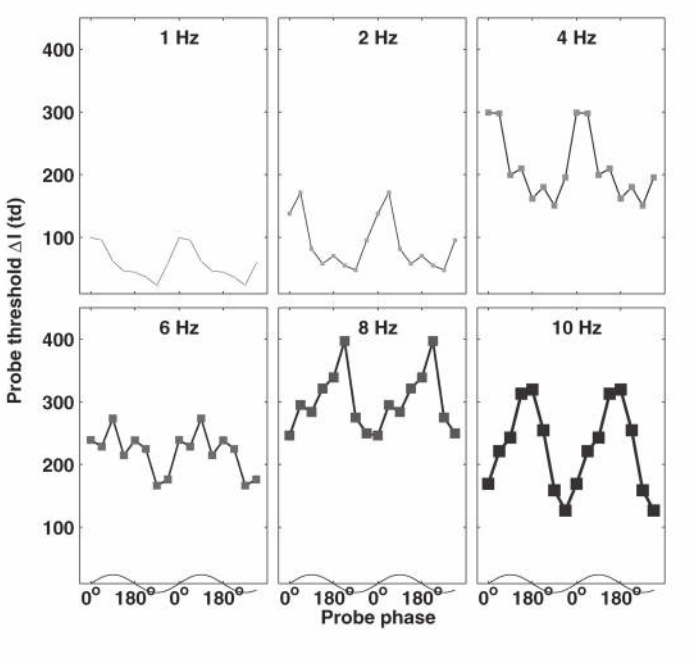

Figure 2.2. Experimental results from the probed-sinewave paradigm, plotted as probe threshold ΔI versus phase. Two cycles of the background are shown for help in displaying the patterns in the results but the points in the second cycle are identical to those in the first cycle. Experimental data from observer JG in Figure 4 of Hood *et al.* (1997).

model of Snippe, Poot, and van Hateren (2000, 2004) shown in Figure 2.4. The Snippe *et al.* model with its three general kinds of processes easily handles the empirical result we have been talking about, that is, the general elevation of probe threshold as background flicker frequency increases from low to middling. The model does so by adding a contrast-gain-control process to the previously suggested subtractive and divisive stages of luminance-controlled processes (light adaptation). See the figure legend for some more details. (The actual processes themselves are described precisely in Snippe *et al.*, 2000.)

The contrast-gain-control process in the model of Figure 2.4 presumably acts before any stage at which there is substantial binocular combination, and therefore its physiological substrate is likely to be the in the retina or LGN. This presumption comes from a further empirical result: In the probed-sinewave paradigm, most of the general elevation with increasing flicker frequency does not show interocular transfer (Wolfson and Graham, 2001b).