Jones, M. G., Bull, D., Canagarajah, C. N., Nikolov, S. G., & Gilchrist, I. D. (2003). A system for gaze-contingent image analysis and multi-sensorial image display. 749 - 756.

A System for Gaze-Contingent Image Analysis and Multi-Sensorial Image Display

S. G. Nikolov[1], I. D. Gilchrist[2],
D. R. Bull[1], C. N. Canagarajah[1]
[1]Centre for Communications Research
[2]Department of Experimental Psychology
University of Bristol, UK
Stavri.Nikolov@bristol.ac.uk

M. G. Jones
School of Information Technology
Griffith University
Gold Coast
Australia
Michael.Jones@gu.edu.au

Abstract – *A novel system for gaze-contingent image analysis and multi-sensorial image display is described. The observer's scanpaths are recorded while viewing and analysing 2-D or 3-D (volumetric) images. A region-of-interest (ROI) centred around the current fixation point is simultaneously subjected to real-time image analysis algorithms to compute various image features, e.g. edges, textures (2-D) or surfaces and volumetric texture (3-D). This feature information is fed back to the observer using multiple channels, i.e. in visual (replacing the ROI by a visually modified ROI), auditory (generating an auditory display of a computed feature) and tactile (generating a tactile representation of a computed feature) manner. Thus, the observer can use several of his senses to perceive information from the image which may be otherwise hidden to his eyes, e.g. targets or patterns which are very difficult or impossible to detect. The human brain then fuses all the information from the multi-sensorial display. The moment the eyes make a saccade to a new fixation location, the same process is applied to the new ROI centred around it. In this way the observer receives information from the local real-time image analysis around the point of gaze, hence the term gaze-contingent image analysis. The new system is profiled and several example applications are discussed.*

Keywords: gaze-contingent, image analysis, multi-sensorial, display, information fusion, eye-tracking

1 Introduction

The human visual system by far still surpasses any computer vision system in terms of performance, speed, accuracy and richness of the processed information. In a way, the human visual system is still the ultimate image processing machine. There are, however, many cases where a human observer is presented with very complex imagery, and successfully carrying out a particular visual task, such as target detection, navigation or classification, is tremendously difficult or even impossible without additional aid. Such assistance could be provided by systems which enhance our vision by, for example, offering additional information either derived by means of computer analysis of the same image or coming from other image sources or modalities. Vision, however, is only one of the senses we use to gather information about the world around us and to interact with it. The human brain is perhaps the best example of a data fusion system that integrates multi-sensorial information, i.e. sight, sound, smell, taste, and touch data, and thus makes inferences regarding the surrounding environment. Humans routinely perform simple and complex tasks in which ambiguous auditory, visual, tactile, etc. data is combined in order to support accurate perception and decision making. In contrast, automated approaches for processing multi-modal data sources still lag far behind [1]. This was one of the primary motivations for designing the proposed system for gaze-contingent image analysis and multi-sensorial image display with the human being the central part of this system.

2 Muti-Sensorial Systems and Information Fusion

How exactly the brain combines information from different sensory modalities is still one of the great mysteries of human perception. Several recent projects around the world are trying to find answers to this difficult problem by studying how people process and fuse information from different senses. Until recently, most of the techniques needed to study two or more senses at once were not available. Hence little research has been done so far to investigate how the brain

creates a coherent perception of the world from the diverse information it receives.

As already discussed, humans use multiple sources of sensory information to estimate environmental properties. For example, the eyes and hands both provide relevant information about an object's shape. The eyes estimate shape using information about binocular disparity, perspective projection, etc., while the hands supply haptic shape information by means of tactile and proprioceptive cues [2]. Ernst and Bülthoff at the Max Planck Institute for Biological Cybernetics in Tübingen, Germany, and Martin Banks at the University of California at Berkeley, have been studying visual-haptic systems and interfaces for a number of years. They recently discovered that the human brain fuses visual and haptic information in a statistically optimal fashion. In [3] they proposed a general principle, which minimises variance in the final estimate and determines the degree to which vision or haptics dominates. This principle is realised by using maximum-likelihood estimation to combine the inputs. Thus, the authors suggest that the nervous system seems to combine visual and haptic information in a fashion that is similar to a maximum-likelihood integrator [3]. Visual dominance occurs when the variance associated with visual estimation is lower than that associated with haptic estimation.

Some very recent studies conducted at the University of California at Berkeley, found that when the human brain is presented with conflicting information about an object from different senses, it finds a remarkably efficient way to sort out the discrepancies. Hillis and colleagues [2] found that when sensory cues from the hands and eyes differ from one another, the brain effectively 'splits' the difference to produce a single mental image. This middle ground can be considered a "weighted average" because in any given individual, one sense may have more influence than the other. When the discrepancy is too large, however, the brain reverts to information from a single cue to make a judgement about what is true. Thus, combining information across cues can improve estimation of object properties but may come at a cost: loss of single-cue information. The results from the experiments described in [2] show that single-cue information is indeed lost when cues from within the same sensory modality (disparity and texture gradients in vision) are combined, but not when different modalities (vision and haptics) are combined.

Lynne Bernstein and her team at the House Ear Institute, USA, are conducting experiments to study whether speech is processed by the brain in the same manner as it processes other types of stimuli. This project uses both brain and behavioural methods to investigate how the speech perceiving brain combines auditory and visual speech under noisy conditions. The main aim of the project is to try to explain the fact that being able to see a talker under noisy conditions dramatically improves the ability to hear that talker's speech. When measured, this effect is equivalent in some cases to almost quadrupling the loudness of the speech signal. A fundamental question is whether this effect occurs because listeners correlate speech information from the talker's lips and face with speech sounds or whether the effect occurs whenever a visual object is paired with speech.

Fisher and Darrell at the MIT Artificial Intelligence have designed and studied multi-modal (audio-video) perceptual user interfaces. In [1] they present an information theoretic approach for fusion of multiple modalities. They also present some empirical results demonstrating audio-video localisation and consistency measures.

3 Gaze-Contingent Displays

The human visual system can only resolve detailed information within a very small area at the centre of vision. Resolution rapidly drops in the visual periphery. This is probably due to the spatial density of cone receptors in the fovea and the retinal ganglion cells in the periphery. Effectively, at any one time our visual system processes information only from a relatively small region centred around the current fixation point. Real-time monitoring of gaze position, using various kinds of eye-tracking devices, permits the introduction of display changes that are contingent upon the spatial or temporal characteristics of eye movements. Such displays, called gaze-contingent displays (GCD), have been described in numerous publications and have been used in various applications, e.g. reading, image and scene perception, and visual search studies (see for example [4, 5, 6, 7]). In GCDs a window centred around the observer's fixation point is modified while the observer moves their eyes around the display. In its classical form, this technique obscures all objects from view except those within the window. In reading research the moving mask and moving window paradigms have proven to be invaluable in determining the chronometric and spatial characteristics of processing written text [8]. Due to technical limitations, gaze-contingent window paradigms have frequently been applied to reading studies, but more rarely to scene perception.

Two types of gaze-contingent image displays which have been proposed in the literature in recent years include gaze-contingent multi-resolution displays (GCMRD) and gaze-contingent multi-modality displays (GCMMD).

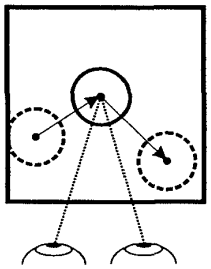

Figure 1: Gaze-contingent display: in most GCD implementations the window information is taken from a 'foreground image' (FI), while the background information comes from a 'background image' (BI).

GCMRDs are GCDs in which image resolution varies with high resolution information being presented at the centre of vision and low resolution information in the periphery. Such displays can lead to great processing and bandwidth saving. An alternative name under which the same idea is described is foveated displays or foveated imaging [9, 10]. The earliest and best known application of GCMRDs is in flight simulators. Some other areas where 2-D GCMRDs have been successfully used include virtual reality, large immersive displays, videoconferencing and tele-operation. A number of studies [11, 12] have been carried out to determine display parameters (window size and border, peripheral degradation) of both imperceptible and perceptible GCMRDs.

GCMMDs [13] are GCDs in which information from one image modality is presented at the centre of vision while information from another (different) modality is presented in the periphery. Some applications of 2-D GCMMDs include image fusion [13] and multi-layered geographical map displays [14]. An example multi-layered road map of Benton County (Washington) is shown in Figure 2 (top). Several layers of the map have been 'switched off' and the result is displayed in Figure 2 (middle). Figure 2 (bottom) presents an GCMMD of the images. 3-D GCMMDs have also been studied in [15] for fusion of volumetric medical images. These have been implemented using a stereo eye-tracker and region-enhanced volume rendering [16].

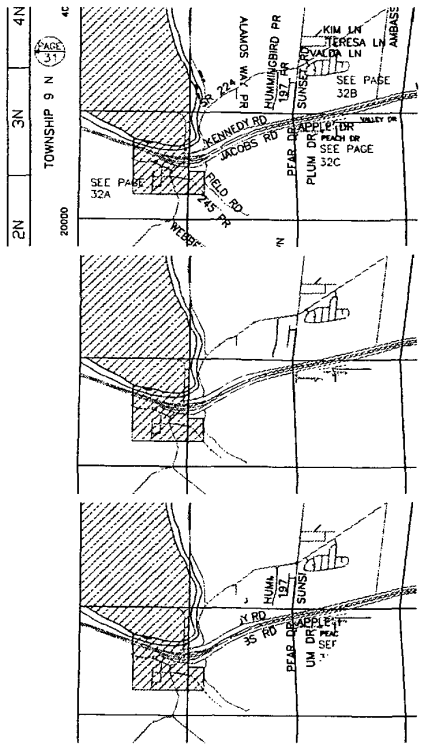

Figure 2: Benton (N. Carywood Road) multi-layered road map: FI - all layers (top); BI - some layers (middle). GCMMD (bottom) of the FI and BI. Image size: 1020×600 pixels. Circular window diameter: 250 pixels. Map courtesy of the Planning/Building Department, Benton County, Washington, USA.

Several hardware and software gaze-contingent display implementations have been proposed in the literature, see for example [8, 7]. Our initial 2-D GCD implementation using an EyeLink I system (from SR Research Ltd.) is described in detail in [13]. A newer implementation using the same eye-tracker but based on texture mapping and OpenGL, which achieves much better performance, is presented in [17]. For more information about the 3-D GCD display we have built around a 3-D stereo eye-tracker refer to [16].

While much research has been done on defining visual ROIs, trying to estimate fixation positions, and correlating the two (see for example [18]) very few studies have been carried out to locally analyse the image content at and around the fixation point. Reinagel and Zador [19] studied natural scene statistics at the centre of gaze and reported that active selection affected the statistics of the stimuli

encountered by the fovea. They found two related effects [19]: (a) that subjects look at image regions that have high spatial contrast; and (b) that in these regions the intensities of nearby image pixels were less correlated with each other than in images selected at random. All the processing they did was off line, i.e. fovea-sized square ROIs ($1^o \times 1^o$ or 23×23 pixels in their experiments), centred at the observer's point of gaze, were recorded to file every 20 ms, and were then processed and analysed afterwards. As far as we are aware no previous experiments have been carried out to locally analyse image content around the fixation point in real time.

4 System for Gaze-Contingent Image Analysis

The new system we have developed is very different from the GCDs described above. It implements a multi-sensorial GCD, where information from a ROI around the current fixation point is processed and analysed by a computer in real time and then the output is send to several of the senses of the observer, thus enhancing and complementing their visual experience and perception. The new system comprises a number of modules: (a) an image display (ID) subsystem; (b) a gaze-tracking (GT) subsystem; (c) an image processing (IP) subsystem; (d) an image analysis (IA) subsystem; and (e) an image sensation and perception (ISP) subsystem. Figure 3 illustrates the general architecture of our system. Below we will describe the role of each subsystem. The image display (ID) subsystem loads a 2-D image from a file and displays it on the screen. The gaze-tracking (GT) subsystem provides real-time information about the observer's fixation point. This subsystem uses an eye-tracking device. The methodology used for measuring eye movements is irrelevant, i.e. any such can be used, e.g. video-oculography or video-based combined pupil and corneal reflection. However, the eye-tracker which is used must provide fast sampling rates (usually around 250Hz or better) and high spatial resolution (better than 0.2^o in most cases, although for some applications high resolution may not be needed). The image processing (IP) subsystem, obtains in real-time a circular region-of-interest (ROI), centred around the current fixation point, from the GT subsystem and processes it simultaneously by n filters (or filter sequences) to produce n filtered ROIs $(R_1, R_2, ...R_n)$. Then, the image analysis (IA) subsystem derives a set of features $(F_1, F_2, ...F_n)$ from each one of the filtered ROIs $(R_1, R_2, ...R_n)$. The image sensation and perception (ISP) subsystem then uses the computed features $(F_1, F_2, ...F_n)$ or the filtered ROIs $(R_1, R_2, ...R_n)$ to generate m sensation maps $(S_1, S_2, ...S_m)$. These sensation maps are presented to the observer simultaneously using multiple channels, e.g. by visualisation (replacing the ROI by a filtered ROI), sonification (generating an auditory display of a computed feature) and tactile feedback (generating a tactile representation of a computed feature). Thus a multi-sensorial image display is constructed.

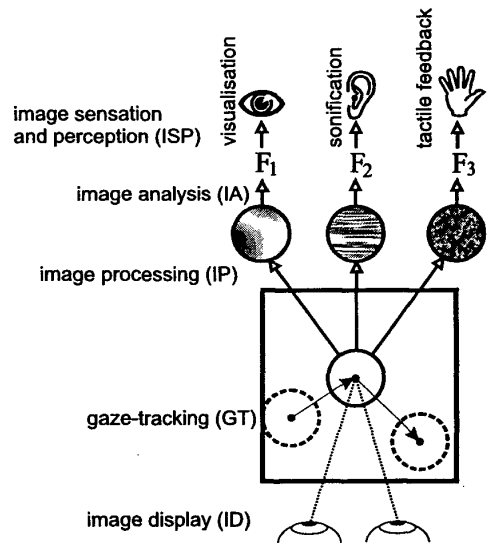

Figure 3: Gaze-contingent 2-D image analysis and multi-sensorial image display.

The new system for gaze-contingent analysis of 3-D (volumetric) images which we constructed has very similar architecture to the 2-D system described above. It consists of the same number and type of subsystems. Figure 4 shows the general architecture of the proposed 3-D system. The main differences between the 3-D and 2-D systems are explained below. The ID subsystem in this case loads a 3-D volumetric image from a file and displays it on the screen, e.g. by using direct volume rendering. The GT subsystem provides real-time information about the observer's point of gaze in 3-D space. The eye-tracker must be capable of computing also the depth of the fixation point. The eye-tracker used in our system is a stereo eye-tracker which provides fast sampling rates and high spatial resolution (see [16] for more details about this system). The IP subsystem takes a spherical ROI centred around the current fixation point and processes it simultaneously by n filters (or filter sequences) to produce n filtered 3-D ROIs

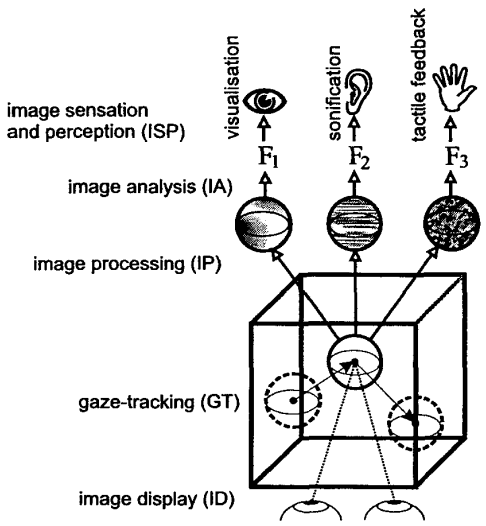

Figure 4: Gaze-contingent 3-D (volumetric) image analysis and multi-sensorial image display.

$(R_1, R_2, ...R_n)$. The IA subsystem then derives a set of features $(F_1, F_2, ...F_n)$ from each one of the filtered ROIs. Following that, the ISP subsystem generates m sensation maps $(S_1, S_2, ...S_m)$ from the computed features $(F_1, F_2, ...F_n)$ or the filtered ROIs $(R_1, R_2, ...R_n)$. Similar to the 2-D case, these sensation maps are sensed and perceived by the observer at the same time using multiple input channels, e.g. by visual, audio and tactile feedback.

5 System Profiling

There are various definitions of 'real-time systems' amongst different communities, such as researchers, software developers, control engineers, etc. The Oxford Dictionary of Computing [20] gives us the following definition of a real-time system: *'A system in which the time at which the output is produced is significant. This is usually because the input corresponds to some movement in the physical world, and the output has to relate to that same movement. The lag (delay) from the input time to output time must be sufficiently small for acceptable timeliness'.* Another definition that is provided in the Journal of Systems and Control Engineering [21] is *'Real-Times Systems are those which must produce correct responses within a definite time limit. Should computer responses exceed these time bounds then performance degradation and/or malfunctions results'.* In the context of this research we will define *a real-time system to be one where the results from the image analysis done locally around the fixation point are computed and channelled simultaneously to a number of our senses with very small delay so that this information is perceived and processed by the observer as coming from a source around the point of gaze.* The main question then is how much time do we have for gaze-tracking, display, image processing and analysis, image sensation and perception, for a system to have such real-time performance, and what can be done in that time. Below are some general considerations about the different subsystems and their performance which are valid for a wide range of applications.

1. ID - The time needed to display an image or a volume, or to update a ROI on the screen, very much depends on the system architecture, and the graphics card and the display in particular.

2. GT - The relationship between saccade duration T (in ms) and saccade amplitude A (in degrees), is well described by the expression $T = 2.2A + 21$ [22]. So a typical 10° saccade will take about 40ms. Following the saccade there is a period during which the eye lens is distorted as a result of the movement. The duration of this distortion varies with saccade size and individual but typically can be of the order of 5-20ms. During this period vision will be at least partially compromised. Fixation durations, when the eye is still, last for somewhere between 100 and 300ms [23] with typical fixation duration of about 250ms. The time span of a fixation and the time used by the IP and IA, and ISP subsystems are illustrated in Figure 5. Generally, at most 50ms is available for image analysis, with another 50-100ms for visual, auditory or tactile feedback (multi-sensorial display).

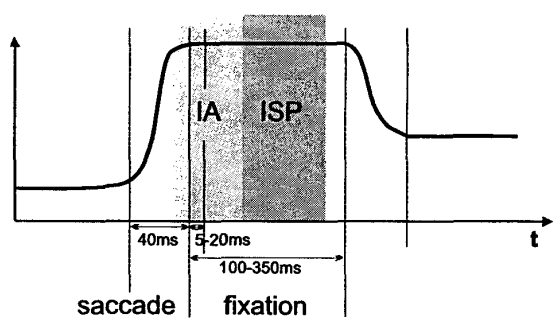

Figure 5: Fixation time span and time available to the IA and ISP subsystems.

3. IP and IA - The time needed for different image analysis operations (2-D and 3-D) strongly de-

pends on the computer architecture and the computational complexity of the operators used. A dual-processor Intel Pentium III or IV PC, for example, is generally capable of analysing in software small 2-D ROIs ($0.5^o \times 0.5^o$ or 13×13 pixels) for some simple operators like the Roberts and Sobel edge detectors with almost no delay. In an OpenGL implementation of a GCD it is possible to achieve real-time image processing and analysis using OpenGL extensions such as the *ARB_IMAGING* imaging subset included in OpenGL version 1.2 [24] or later. Of course in this case the user is limited to the set of filters and operators being supported.

4. ISP - The time needed to build sensation maps depends upon the complexity of the mapping between the (image) feature domain and the target domain (audio, tactile, etc.). Sufficient time is also needed for the observer to be able to perceive the information from the image analysis fed back along one or multiple channels.

Recent results reported by Hillis and colleagues [2] show that at least in some cases single-cue information is lost when cues from within the same sensory modality are fused, but not when different modalities, e.g. vision and haptics, are combined. It is possible then, that the brain keeps information from different sensory cues separate, at least initially. Another important result from their study is that when discrepancies in the information obtained by different sensors become too large the brain turns to only one sense, probably depending upon which one seems more accurate. In the context of our system, this means that in general utilising a multi-sensorial display to show image information, derived by means of gaze-contingent image analysis, could be expected to provide the brain with a richer set of information than using visual cues alone. It also opens the question of what can be considered to be conflicting information in this framework or when such information coming from different senses is perceived by the brain as contradictory. In some cases that could be obvious, e.g. in the second example in the next section if we see an object that is horizontal but hear a sound from the system that tells us it is vertical. However, in many cases where more complex mappings are used this contradiction could be very subtle. Hence, it is very important in the future to study when and to what extent the brain perceives multi-sensorial information as contradictory in the context of particular applications and tasks. This could be used as a measure of 'goodness' of the sensation maps being constructed.

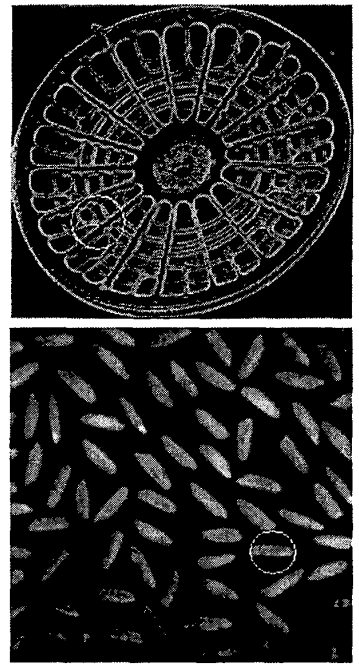

Figure 6: Example demos using gaze-contingent image analysis: edgedness demo (top); orientation demo (bottom). Diatom image (top) from the ADIAC project (University of Algarve, Portugal). Rice grains image (bottom) courtesy of The Mathworks Inc.

6 Examples and Applications

Several demonstrations have been developed which embody the ideas described above. Two simple examples include a demo which measures the 'edgedness', i.e. the amount of edges (using Roberts or Sobel edge detectors), of the area around the fixation and produces a short 'beep' if this amount is above a certain threshold. This information can be useful in terms of finding edge-rich regions (Figure 6 (top)). Computing various texture properties can be done in a similar way. Another demo determines the orientation of certain objects in an image and if they are within a predefined range again 'beeps' when the observer's gaze is over such an object. For instance, such a system will warn the observer when they are looking at 'horizontal' (in this case defined as having their main axis in the range $[-0.5^o, +0.5^o]$)

rice grains in the image in Figure 6 (bottom). Various kinds of image analysis can be performed depending on the application. Some applications which could perhaps benefit most from the use of gaze-contingent image analysis are: (a) analysis and understanding of complex imagery, e.g. 2-D remote sensing or medical images, natural scenes, fused images or multi-modality displays, 3-D medical, seismic, or sonar volumes; (b) target detection and pattern or object recognition in such images, e.g. camouflaged target detection in remote sensing images, crack detection in industrial inspection, detection of irregularities in textures; and (c) prompting and alerting systems.

7 Conclusion

A new system for real-time gaze-contingent image analysis has been described. The essential properties of such a system include: (a) the ability to record eye movements using eye-trackers with fast refresh rates and high accuracy; (b) a fast display system which can update ROIs on the screen with minimal delay and flicker; (c) a computer which can perform various image analysis tasks in very short periods of time (usually less than 50ms), and perhaps in parallel, and can compute speedily various sensation maps; and (d) the ability to feed back in real time the information from the sensation maps using several of the observer's senses. All these properties of the system very much mimic some of the characteristics of the way we humans sense, perceive and process information from the surrounding world. However, in many cases where we may fail in difficult visual analysis and detection tasks, a system for gaze-contingent image analysis using a multi-sensorial display could be a valuable prompting and alerting tool that can hopefully complement and enhance our otherwise wonderful vision. Further studies are needed to investigate which tasks can be aided by only local (gaze-contingent) image analysis and to what degree, and also how we combine this local information to build high-level semantic models of the world we view. The way our brain fuses multi-sensorial information is still very much an open issue which will undoubtedly be investigated in various applications in the years to come, with an increasing number of multi-modal interfaces and systems being designed and used in everyday life.

8 Acknowledgements

Financial support from the Royal Society (UK) is gratefully acknowledged.

References

[1] John Fisher and Trevor Darrell. Signal level fusion for multimodal perceptual user interface. In *Workshop on Perceptive User Interfaces.* ACM Digital Library, November 2001.

[2] J. M. Hillis, M. O. Ernst, M. S. Banks, and M. S. Landy. Combining sensory information: Mandatory fusion within, but not between, senses. *Science*, 22:1627–1630, November 2002.

[3] M. O. Ernst and M. S. Banks. Humans integrate visual and haptic information in a statistically optimal fashion. *Nature*, 415:429–433, January 2002.

[4] G. W. McConkie and K. Rayner. The span of the effective stimulus during a fixation in reading. *Perception and Psychophysics*, 17:578–586, 1975.

[5] S. Saida and M. Ikeda. Useful visual field size for pattern perception. *Perception and Psychophysics*, 25:119–125, 1979.

[6] K. Rayner. Eye movements in reading and information processing: 20 years of research. *Psychological Bulletin*, 124:372–422, 1998.

[7] M. Pomplun, E. M. Reingold, and J. Shen. Investigating the visual span in comparative search: the effects of task difficulty and divided attention. *Cognition*, 81:B57–B67, 2001.

[8] P. M. J. van Diepen et al. On-line control of moving masks and windows on a complex background using the ATVista Videographics Adapter. *Behavior Research Methods, Instruments and Computers*, 26:454–460, 1994.

[9] C. F. R. Weiman. Video compression via log polar mapping. In *Real Time Image Processing II*, volume 1295, pages 266–277. SPIE, 1990.

[10] P. T. Kortum and W. S. Geisler. Implementation of a foveated image-coding system for bandwidth reduction of video images. In B. Rogowitz and J. Allebach, editors, *Human Vision and Electronic Imaging*, volume 2657, pages 350–360. SPIE, 1996.

[11] L. C. Loschky and G. W. McConkie. User performance with gaze contingent displays. In *Proceedings of the Eye Tracking Research and Applications Symposium 2000*, pages 97–103. ACM SIGGRAPH, 2000.

[12] E. M. Reingold and L. C. Loschky. Reduced saliency of peripheral targets in gaze-contingent multi-resolution displays: Blended versus sharp boundary windows. In *Proceedings of the Eye*

Tracking Research and Applications Symposium 2002, pages 89–94. ACM SIGGRAPH, 2002.

[13] S. G. Nikolov, M. G. Jones, I. D. Gilchrist, D. R. Bull, and C. N. Canagarajah. Multi-modality gaze-contingent displays for image fusion. In *Proceedings of the 5th International Conference on Information Fusion (Fusion 2002), Annapolis, MD, USA*, pages 1213–1220. ISIF, 2002.

[14] S. G. Nikolov, I. D. Gilchrist, and D. R. Bull. Gaze-contingent multi-modality displays of multi-layered geographical maps. In *Proceeedings of the Fifth International Conference on Numerical Methods and Applications (NM&A02), Symposium on Numerical Methods for Sensor Data Processing, Borovetz, Bulgaria*, volume 2542, pages 325–332. Lecture Notes in Computer Science, Springer-Verlag, 2003.

[15] S. G. Nikolov, M. G. Jones, D. R. Bull, C. N. Canagarajah, M. Halliwell, and P. N. T. Wells. Focus+context visualisation for image fusion. In *4th International Conference on Information Fusion (Fusion 2001), Montreal, Canada, 7-10 August*, volume I, pages WeC3–3 – WeC3–9. ISIF, 2001.

[16] M. G. Jones and S. G. Nikolov. Volume visualisation via region-enhancement around an observer's fixation point. In *MEDSIP 2000 (International Conference on Advances in Medical Signal and Information Processing), Bristol, UK, 4-6 September, 2000*, number 476, pages 305–312. IEE, 2000.

[17] Tim Newman. *Gaze-contingent displays using texture mapping, MEng Thesis.* University of Bristol, UK, Bristol, UK, 2003.

[18] C. M. Privitera and L. W. Stark. Algorithms for defining visual regions-of-interest: comparison with eye fixations. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 22(9):970–982, September 2000.

[19] P. Reinagel and A. M. Zador. Natural scene statistics at the centre of gaze. *Network: Comput. Neural. Syst.*, 10:341–350, 1999.

[20] V. Illingworth and I. Pyle, editors. *Oxford Dictionary of Computing.* Oxford Paperbacks, 1996.

[21] C. R. Burrows, editor. *Journal of Systems and Control Engineering.* Professional Engineering Publishing.

[22] R. H. S. Carpenter. *Movements of the eyes. (2nd edn).* Pion Press, London, 1988.

[23] R. H. S. Carpenter. Oculomotor procrastination. In *Eye Movements, Cognition and Visual Perception.*

[24] M. Segal and K. Akeley. The OpenGL graphics system: A specification (version 1.2.1). Technical report, Silicon Graphics, Inc., Mountain View, CA, October 1998.