CONTENT-PRIORITISED VIDEO CODING FOR BRITISH SIGN LANGUAGE COMMUNICATION

LAURA JOY MUIR

A thesis submitted in partial fulfilment of the
requirements of
The Robert Gordon University
for the degree of Doctor of Philosophy

October 2007

Abstract

Video communication of British Sign Language (BSL) is important for remote interpersonal communication and for the equal provision of services for deaf people. However, the use of video telephony and video conferencing applications for BSL communication is limited by inadequate video quality.

BSL is a highly structured, linguistically complete, natural language system that expresses vocabulary and grammar visually and spatially using a complex combination of facial expressions (such as eyebrow movements, eye blinks and mouth/lip shapes), hand gestures, body movements and finger-spelling that change in space and time. Accurate natural BSL communication places specific demands on visual media applications which must compress video image data for efficient transmission.

Current video compression schemes apply methods to reduce statistical redundancy and perceptual irrelevance in video image data based on a general model of Human Visual System (HVS) sensitivities.

This thesis presents novel video image coding methods developed to achieve the conflicting requirements for high image quality and efficient coding. Novel methods of prioritising visually important video image content for optimised video coding are developed to exploit the HVS spatial and temporal response mechanisms of BSL users (determined by Eye Movement Tracking) and the characteristics of BSL video image content. The methods implement an accurate model of HVS foveation, applied in the spatial and temporal domains, at the pre-processing stage of a current standard-based system (H.264). Comparison of the performance of the developed and standard coding systems, using methods of video quality evaluation developed for this thesis, demonstrates improved perceived quality at low bit rates.

BSL users, broadcasters and service providers benefit from the perception of high quality video over a range of available transmission bandwidths. The research community benefits from a new approach to video coding optimisation and better understanding of the communication needs of deaf people.

Acknowledgements

I am grateful to The Robert Gordon University for giving me the opportunity to embark on a part-time PhD within a year of taking up my lecturing appointment. I wish to acknowledge the contribution of all those who inspired, guided and supported me from inception to completion.

In particular, I wish to thank those individuals who contributed to the development of my knowledge and understanding of sign language, video communication systems and the importance of human factors in the design of systems for people. Lillian Lawson, Director of the Scottish Council on Deafness, provided the motivation for this research. Professor Iain Richardson gave his expert guidance and asked the right questions to help me develop and adduce my ideas. Professor Richardson and my colleagues at the School of Engineering and the Centre for Video Communications; Dr Tony Miller, Dr Yafan Zhao and Dr Sampath Kannangara, proffered highly valued encouragement, advice, support and friendship.

The research conducted for this thesis would not have been possible without the support and participation of the deaf community in Aberdeen. I am most grateful to Edith Ewen, Honorary Chairperson of the Aberdeen Deaf Social and Sports Club, for organising volunteers for the experimental work and to all the very willing members of the Club who participated in the research. I would like to thank Lisa Davidson, Kathleen Cameron and Michael Tocher for signing in the video sequences used in the experimental work. My special thanks go to Jim Hunter, an enthusiastic and indefatigable supporter of the deaf community, for his BSL interpreting skills. I am immensely grateful to Jim, his wife Norma and the profoundly deaf people who participated in this research for giving me an insight into the needs, aspirations and contributions of deaf people. I would also like to thank Isobel Slessor and Mags Christie for their expert BSL tuition and their patience.

Contents

List of Figures

List of Tables

Abbreviations and Acronyms

ACR	Absolute Category Rating
ADSSC	Aberdeen Deaf Social and Sports Club
ASCII	American Standard Code for Information Interchange
ASL	American Sign Language
Auslan	Australian Sign Language
AVC	Advanced Video Coding
BAM	Binary Acceptability Method
BBC	British Broadcasting Corporation
BDA	British Deaf Association
BR	Bit Rate
BSL	British Sign Language
CA	Communications Assistant
CACDP	Council for the Advancement of Communication with Deaf People
CCITT	Comité Consultatif International Téléphonique et Télégraphique (now ITU)
CIF	Common Intermediate Format
CNS	Central Nervous System
CODEC	enCOder/DECoder
CSD	Communication Service for the Deaf (USA)
CSIRO	Commonwealth Scientific and Industrial Research Organisation
CT	Contrast Threshold
CTF	Contrast Threshold Function
CT_0	Minimum Contrast Threshold
DCR	Degradation Category Rating
DCT	Discrete Cosine Transform
DSCQS	Double Stimulus Continuous Quality Scale
DSISM	Double Stimulus Impairment Scale Method
DSM	Double Stimulus Method
DWT	Discrete Wavelet Transform
EMT	Eye Movement Tracking
ERP	Event-Related Potentials

ESRC	Economic and Social Research Council
FDCT	Forward Discrete Cosine Transform
fMRI	functional Magnetic Resonance Imaging
FRExt	Fidelity Range Extensions
FS	Filter Strength
FSVC	Foveation Scalable Video Coding
FWTF	Foveation-Weighted Temporal Filter
HCI	Human-Computer Interaction
HDTV	High Definition Television
HVS	Human Visual System
IDCT	Inverse Discrete Cosine Transform
IEE	Institute of Electrical Engineering (now IET)
IEEE	Institute of Electrical and Electronic Engineering
IET	Institute of Engineering and Technology
IP	Internet Protocol
IR	Infra Red
ISDN	Integrated Services Digital Network
ISL	Irish Sign Language
ISO/IEC	International Standards Organisation/International Electrotechnical Commission
ITU	International Telecommunications Union
ITU-T	International Telecommunications Union - Telecommunication Standardisation Sector
JPEG	Joint Picture Expert Group
JVT	Joint Video Team
kbps	Kilobits per second
LED	Light Emitting Diode
MOS	Mean Opinion Score
Mbps	Megabits per second
MPEG	Motion Picture Expert Group
MSE	Mean Squared Error
PC	Pair Comparison
PCI	Peripheral Component Interconnect

PSNR	Peak Signal-To-Noise Ratio
PSNR Y	Peak Signal-To-Noise Ratio (Luminance)
QCIF	Quarter Common Intermediate Format
QoS	Quality of Service
QP	Quantisation Parameter
QUASS	QUality ASsessment Slider
RNID	Royal National Institute for the Deaf
ROI	Region-of-Interest or Region-Of-Importance
RTSP	Real Time Streaming Protocol
SAE	Sum of Absolute Errors
SD	Standard Deviation
SIP	Session Initiation Protocol
SPSS	Statistical Package for the Social Sciences
SSCQE	Single Stimulus Continuous Quality Evaluation
SSM	Single Stimulus Method
TRS	Telecommunications Relay Service
TSCR	Task-Specific Category Rating
UFQ	User Feedback Quality
VCEG	Video Coding Expert Group
VO	Video Object
VRI	Video Remote Interpreting
VRS	Video Relay Service

PART ONE: BACKGROUND

Chapter 1: Introduction

Of all the human senses, vision is relied on most heavily for sensory input about the environment (Hendee and Wells, 1997). Understanding the detection, recognition and interpretation of visual information has a tremendous impact on how we format, transmit and use visual information and on the design of information systems. This is particularly important in the design of video communication systems for people who are deaf and rely on visual communication of information using sign language.

This chapter sets the scene for the development of video communication systems optimised for the specific visual requirements of deaf users. The research problem, rationale, hypothesis, aim and objectives for the research are described. The contributions and outputs (Appendix A) and the impact of the research, in terms of the novelty, timeliness and beneficiaries are previewed and an overview of the structure of the remainder of the thesis is described.

1.1 The Research Problem and Rationale

Video communication systems permit the transmission of moving pictures between senders and receivers reciprocally and enable remote person-to-person communication. The more bits that are used to represent video image data, the better the fidelity of the image will be. However, the bit rate generated by video communication systems is limited by the available bandwidth of the communication channel. Video compression is required to reduce the number of bits required for transmission of video information whilst maintaining acceptable quality for the receiver. The design of video compression systems requires a trade-off between the conflicting requirements for efficient coding and high video quality. User perception of quality depends on the type of application (for example: video conferencing, video surveillance and telemedicine applications). Other factors influencing perceived video image quality include the image size, frame rate, spatial resolution and number of light intensity and colour levels. Video communication systems must be optimised to meet the quality requirements of the user.

Video communication systems are of particular benefit to the deaf community who rely on the visual communication of information for interpersonal telecommunication (including video telephony and video conferencing) and access to services (for example, video relay sign language interpretation). The equal provision of services for people who are deaf is a legal requirement for service providers, broadcasters and employers under the UK Disability Discrimination Act (1995) and the Communications Act (2003). There are three main groups of deaf user with specific communication requirements (ITU-T Study Group 16, 1998). The first group includes people who have partial hearing. This group requires audio amplification and speech/lip-reading to support spoken language communication. The second group, people who became profoundly deaf later in life, rely on speech/lip-reading for communication in the spoken language of the country. The third group of deaf users is pre-lingually deaf and is totally reliant on the visual communication of information using sign language. This thesis addresses the video communication needs of this third group of users, specifically people who use British Sign Language (BSL) as their first/preferred language. However, there are some common visual requirements in the three main groups of deaf people and so optimisation of video communication systems for BSL users also benefits deaf people who use speech/lip-reading only.

For BSL users, video communication systems offer greater access to information and better freedom of expression, in the user's first/preferred language, than English language text-based alternatives (McCaul, 1997). However, the use of video telephony and video conferencing applications is limited by the quality and reliability of existing provision. Accurate natural sign language communication places specific demands on visual media applications due to the speed and complex multi-channel mode of delivery, which includes facial expression, finger-spelling, hand gestures and body movements. The International Telecommunication Union – Telecommunication standardisation sector (ITU-T) draft profile (ITU-T Study Group 16, 1998) specified the quality requirements for real-time sign language video communication which included a minimum of Common Intermediate Format (CIF) resolution (that is, 352×288 displayed pixels) and frame rate of at least 25 frames per second. Visual perception of sign language video content requires sufficient spatial and temporal resolution to capture the detailed movements of the signer which convey information to the deaf receiver. Video frame rate in particular has been shown to have a significant effect on the

communication of facial gestures in sign language; especially lip/mouth shapes (Woelders, Frowein, Nielsen, Questa and Sandini, 1997). Reasonable spatial quality and frame rates can be achieved using current video compression Standards such as the latest version of H.263 (ITU-T, 2000) at high bit rates. At bit rates less than two-hundred kilobits per second (kbps), real-time video communication is characterised by low frame rates, small picture size and poor picture quality (Richardson, 2003). Even the improved compression efficiency of the H.264/AVC coding Standard (ITU-T, 2005) is not acceptable for comfortable, accurate sign language communication at low bit rates (discussed in Section 3.3).

Animated signing, particularly the use of avatars, has been developed in an attempt to overcome the problems of communicating real-time BSL video over networks. Signing avatar projects (for example, VisiCast and eSign), developed in partnership with the Royal National Institute for the Deaf (RNID), have been used to provide sign language interpretation of web sites, local government information and television broadcasts (RNID, 2006b). Although the quality of avatars has improved significantly since these projects began in 2000, the deaf community has not embraced this technology as a suitable substitute due to the limited ability of avatars to express the complexity of the language, particularly facial expression (BBC, 2002).

The limitations of current video communication systems and lack of suitable alternatives for deaf people using sign language have driven research which aims to optimise service quality within delivery constraints. General video compression schemes apply methods to reduce statistical redundancy in video image data and perceptual irrelevance based on a general model of Human Visual System (HVS) sensitivities (Chapter 4). These methods are designed to optimise coding efficiency and improve the overall perception of quality across the video scene.

Systems specifically designed for deaf people have been based on segmentation of a designated region (or regions) of ‘interest’ or ‘importance’ in the video scene using object or model based coding schemes and approximate models of the spatial response of the human eye. Region-of-Interest (ROI) schemes were developed in previous research (Schumeyer, Heredia, and Barner, 1997, Saxe and Foulds, 2002) in an attempt to improve the perceived quality of video images by giving coding priority to visually ‘important’ segmented image regions. However, these researchers assumed

that the hands and face must be transmitted at the same (spatial and temporal) resolution and the research lacked application of a suitable subjective quality testing method for end-user evaluation.

Geisler and Perry (1998) demonstrated the potential to exploit the decline in the perception of spatial resolution in the peripheral field of vision in applications where the locus of visual attention was detected and tracked. This led to the development of image foveation techniques which aimed to give coding priority to visually important content at the expense of content observed in peripheral vision. Implementations of this approach in block-based coding schemes (developed after the research approach in this thesis was first published in 2002) were limited to approximate models of foveation and indicated problems with rate control and artefacts at block boundaries (Sheikh, Evans and Bovik, 2003; Agrafiotis, Canagarajah, Bull and Dye, 2003 and Agrafiotis, Canagarajah, Bull, Kyle, Seers and Dye, 2006).

Approximate models of vision, assumptions about visual behaviour, limited consultation with end-users and inappropriate subjective quality testing methodologies in previous research contributed to a lack of progress in the development of video communication systems for the specific visual communication requirements of BSL users. Support for further research in this field from Lillian Lawson, Director of The Scottish Council on Deafness, and the deaf community in Aberdeen provided the motivation for this thesis.

This thesis explores the visual response mechanisms of BSL users and the important information-carrying content of BSL video images in the design and development of a novel video communication system optimised for deaf people. Eye Movement Tracking (EMT) during BSL video communication tasks is conducted to determine the locus of visual attention of BSL users. The location of visual fixations enables the determination of the BSL video image content viewed in high resolution (foveal) vision and in low resolution (peripheral) vision. This identifies the priority (given at the point of gaze) by the Human Visual System (HVS) to regions-of-importance for the BSL communication task. BSL video images are analysed to determine the motion characteristics (optical flows) in the visually prioritised regions-of-importance and in the image regions observed in peripheral vision. Comparison of the video image content which is given priority by the HVS with the coding priority given in a standard video CODEC is made to inform the design of novel methods of spatial and temporal video image coding. The

content-prioritisation methods are developed and applied at the pre-processing stage of video coding and the perception of quality is compared with standard H.264 video compression at fixed bit rates. New methods of video quality assessment are developed to measure the perception of quality for the BSL communication task.

1.2 Research Hypothesis

It is postulated that by giving priority to visually important content, higher video compression ratios can be achieved while maintaining the perception of high quality.

What is visually important depends on the nature and requirements of the task being performed while viewing the visual stimulus. BSL communication is an active complex task with specific visual characteristics.

By studying the visual response to BSL communication, with respect to the nature and characteristics of the task, visually important content can be identified and prioritised in the design of an efficient video communication system. Giving coding priority (in terms of bit allocation) to visually important content, at the expense of less visually important video image regions, reduces bandwidth requirements and improves the perceived quality of BSL communication.

1.3 Aim and Objectives

The aim of this research is to improve the perceived quality of BSL video communication for deaf users. The objectives are:

1) To evaluate the human factors and design constraints affecting the development of video communication systems for deaf people using BSL.
2) To investigate the visual response of deaf viewers to BSL video image content.
3) To develop a novel method(s) of video compression for BSL communication.
4) To develop a methodology for assessing the perceived quality of BSL video communication.
5) To demonstrate improved performance of the developed system(s) compared to standard systems.

1.4 Research Contributions

The multidisciplinary approach to the optimisation of video coding for BSL communication in this thesis establishes a new way of addressing the research problem. This research contributes to the body of knowledge extending across different research disciplines, including video image processing, sign language studies, human-computer interaction, human vision and visual psychology. The main contributions are described in this section and reviewed in the context of methodological and substantive innovations in the conclusion of this thesis (Section 12.2).

1.4.1 New Approach to Video Coding Optimisation

A novel video image content prioritisation approach to optimising video coding for BSL video communication, based on visual response mechanisms determined by Eye Movement Tracking (EMT), is developed in this thesis. The new approach was presented to the video communication and multimedia applications research communities at an Institute of Electrical Engineers (IEE, now known as IET) seminar (Richardson, Zhao and Muir, 2002) in London and at an International ACM (Association for Computing Machinery) conference (Muir and Richardson, 2002) in France. The publication of this world-first research was followed by an invitation to contribute to an ESRC funded seminar at the University of Bristol (Muir and Richardson, 2003) which brought academic researchers, technology providers, and representatives of the deaf community together for knowledge sharing and to direct future research in video communication systems for deaf people. The EMT and video image content prioritisation approach developed in this thesis was also presented in a paper (Muir, Richardson and Leaper, 2003) at an International Picture Coding Symposium in France.

1.4.2 New Methods of Measuring the Perceived Quality of BSL Video Communication

This research identifies the limitations of existing methods of assessing video quality and the lack of user involvement in the development and testing of video communication systems for deaf people in previous research (Chapter 5). A user-centred design approach is critical to the development of a system designed to meet the specific visual requirements of deaf people for BSL communication. New methods

of subjective quality assessment are developed which enable evaluation of user satisfaction, in terms of acceptability for the communication task, rather than overall image fidelity. The methods are designed to apply assessment criteria which can be easily and accurately expressed in BSL, avoid interference with the primary task (BSL communication) and minimise the time, effort and memory requirements from participants. Novel methods, using a binary or five-point task-specific rating scale and procedures based on ITU recommendations for subjective quality assessment of multimedia content, are developed for this thesis (Chapter 9). The Task-Specific Category Rating (TSCR) scale and subjective quality evaluation methodology developed for this thesis, the results obtained from the application of the method and the content-prioritisation approach developed from the research were presented at the Visual Information Engineering conference in Glasgow (Muir, Richardson and Hamilton, 2005).

1.4.3 New Insights into BSL communication

Following publication of the methodological approach in this thesis in 2002 and 2003 (described in Section 1.4.1), further experimental work was conducted to provide a more in-depth study of visual perception mechanisms (Section 6.4). This work was published in an international journal specialising in deaf studies and provided this research community with a new insight into the way deaf people communicate using sign language in the context of the development of video communication systems for deaf people (Muir and Richardson, 2005).

1.4.4 New Methods for Content-Prioritised BSL Video Coding

New methods of prioritising visually important video image content for optimised video coding are developed to exploit the HVS spatial and temporal response mechanisms of BSL users and the characteristics of BSL video image content (Chapter 8). These methods are applied in standard video CODECs (H.263 and H.264[1]). The methods implement an accurate model of HVS foveation, applied in the spatial and temporal

[1] Application in either Standard is acceptable since the method is implemented at the pre-processing stage.

domains, which produces efficient video image compression and the perception of quality for the BSL communication task. Comparison of the performance of the developed and standard coding systems, using new methods of video quality evaluation developed for this thesis, demonstrates improved subjective quality at low bit rates (Chapter 10) which is a significant contribution to research in this field. The content-prioritised video coding methods for BSL communication developed and tested in this thesis can be applied without modification of existing video coding standards which is an advantage in the development of systems for the small-sized commercial market of BSL users.

1.4.5 Perceptually Optimised Video Compression

The outputs from this research contributed to a Proof of Concept project funded by Scottish Enterprise in 2006. The project aims to develop perceptually optimised video for a wide range of applications (including video conferencing, surveillance and sign language) based on the methods developed in this thesis and the work of other researchers at the Centre for Video Communications at The Robert Gordon University. The major contribution of this thesis to the Proof of Concept project is the novel approach to optimisation based on perceptual mechanisms.

1.5 Research Impact

This research is important at a time when broadcasters and service providers are competing to meet the increasing demand for high quality multimedia services stimulated by the wide availability of broadband services in the UK and the falling cost and rising capacity of storage and transmission. At low and high transmission bandwidths, efficient video compression is required to achieve significant performance advantages and improved perceptual quality within the constraints of increasing user demand and ‘premium’ commercial tariffs set for higher bandwidth limits. Improved coding efficiency enables the coding of more video channels or higher quality video representations within existing digital transmission capacities. Video telephony applications for sign language users (including the full range of vital services such as the development of Video Relay Services) benefit from robust, efficient video coding optimised for the specific needs of the user.

The equal provision of services for people who are deaf is also a legal requirement for broadcasters and service providers who must meet the targets for delivery set by the Office of Communications (Ofcom), the telecommunications regulatory body in the UK.

The development of new services in the future, for example ‘closed signing’ on digital television channels will also benefit from more efficient coding of BSL video images. Current ‘open signing’ systems have a signer translating media content in one corner of the television screen image for a particular broadcast programme. The advantage of this method is that no new technology or extra bandwidth is required, as the signer is part of the main image. However, the wider audience tends to find it intrusive and so open signed programmes are usually limited in quantity and are generally broadcast outside peak viewing times. An alternative approach is to develop technology for closed signing which viewers can switch on or off by pressing a button on their television remote control unit. This would enable better service provision to deaf viewers without intruding on the content provided to other customers. Broadcasters could deliver closed signing by making a secondary parallel stream of signing available for the main programme. When the service is made available, the secondary picture would be shown together with the main broadcast. However, this would require more bandwidth (which is a scarce resource in broadcast channels) and so to make closed signing practically and economically viable, the signing stream must be as small in size as possible. This requires a high degree of image compression and the perception of high quality for the deaf viewer. Delivery of the service would also require functions in television equipment to recognise and display the signing and agreements between broadcasters to provide the secondary streams in a standard format.

The novelty of this thesis is the optimisation of video communication systems based on visual perception mechanisms. This has led to the development of content-prioritised video coding methods for BSL communication based on the spatial and temporal visual responses of deaf people and analysis of the content of the visual stimulus. The development of a new method of measuring the perceived quality of BSL video communication has enabled the perceptual gain of the developed system compared to standard video compression methods to be determined in terms of user acceptability. The application of this method has demonstrated improved performance of the developed system compared to standard video coding methods. The conflicting

requirements for efficient compression and perception of quality for the BSL communication task are achieved in the novel approach developed in this thesis.

The beneficiaries of the developed methods are the BSL users, broadcasters and service providers who profit from high quality video over a range of available transmission bandwidths. The research community benefits from a new approach to video coding optimisation and better understanding of the communication needs of this user group.

1.6 Structure of the Thesis

The thesis is organised in three main parts. The content of the chapters in each part is described in this section and illustrated with reference to the research objectives of the thesis in Figure 1.1.

Part One (Chapters 1 to 5) provides the research context. It explores aspects of the Human Visual System, BSL communication, video communication and video quality assessment relevant to the scope of this research. The human factors and design constraints which have an impact on the development of video communication systems for deaf people (Research Objective 1) are evaluated.

Chapter 2 describes the Human Visual System (HVS) and the visual processing capabilities and limitations and the perceptual mechanisms by which the HVS samples the visual scene and obtains information about features and objects (including faces). It identifies the strong association between eye movements and visual attention for complex visual tasks such as BSL communication and provides the rationale for a study of eye movements of deaf people as a means of identifying visually important image content in BSL video. A review of the spatial and temporal response (including the visual processing of motion) identifies the limitations of the HVS and the enhanced capabilities of deaf viewers which can be exploited in the design of BSL video communication systems. Chapter 2 concludes by defining a model of visual perception for this thesis. It identifies the oculomotor, visual, cognitive, contextual and human factors and phenomena which influence the perception of BSL video communication.

Chapter 3 explores the nature of the visual task, BSL communication. It identifies the range of available video telephony options and the visual quality requirements and visual behaviour mechanisms for BSL communication. This chapter describes the diversity, complexity and visual representation of BSL and addresses the misconceptions about sign languages which have prevailed in previous research efforts to design systems for deaf people. It describes the range of video services and developments in telecommunications that have led to increasing demand for video communication for deaf people for interpersonal communication and access to services. The barriers to service entry and to effective communication, identified in Chapter 3, provide the rationale for optimisation of systems which meet the specific task requirements, quality of service requirements and visual behaviour mechanisms of BSL users.

Current video compression standards and methods of optimisation are evaluated in Chapter 4. This Chapter reviews the method and purpose of efficient digital video data compression for delivering perceived quality within the constraints of increasing user demand and limited transmission bandwidth. International video coding standards are explored to identify options for improving efficiency and developing optimised video communication systems. 'Lossy' compression methods (Section 4.1.3), necessary to achieve the required compression ratios for video communication systems, are explored as a means of removing subjective redundancy (based on knowledge of the sensitivities of the HVS) without significant impact on the perception of visual quality by the user. In the design of lossy systems for BSL video communication, where the loss of visual information has an impact on the understanding of the language, optimisation must prioritise visually important content in the coding scheme. Chapter 4 concludes by reviewing the options for ROI (Region-of-Importance) priority coding and the limitations of methods developed in previous research. The review and evaluation of video coding techniques informs the design, development and application of new methods of content-prioritised video coding in the experimental work in Part Two (Chapter 8) of this thesis.

An appropriate method of assessing video image quality is identified as being critical for optimising quality and bit rate, particularly in a system designed to meet the specific communication needs of BSL users. Chapter 5 evaluates video quality assessment

methods and identifies important criteria for the design of appropriate methods of end-user evaluation of video quality for BSL communication. Automatic 'objective' image quality metrics and subjective video quality methods of assessment by end-users are reviewed. The limitations of standard approaches and the alternative methods developed in previous research are discussed. This chapter informs the design, development and application of new methods of subjective video quality assessment in the experimental work in Part Two (Chapters 9 and 10) of this thesis.

Part Two (Chapters 6 to 10) contains the methods, results and discussion of the experimental work conducted in the thesis. It investigates the visual response of deaf viewers to BSL video content (Research Objective 2). The development of optimised video coding systems (Research Objective 3) and methods of video quality assessment for BSL video communication (Research Objective 4) are described. Improved performance (in terms of perceived quality) compared to a standard system at fixed bit rates (Research Objective 5) is demonstrated.

Chapter 6 explores the visual response to BSL video content in EMT studies of deaf viewers. The identification of image content viewed in 'clear' foveal vision and the information gathered from peripheral vision is important in the development of a video communication system that works optimally within the capabilities and limitations of human vision. Chapter 6 explores the methodological issues raised by previous researchers using EMT. It describes the application of EMT to test the responses of deaf viewers to a wide range of BSL movements and gestures in BSL communication tasks and investigates viewing patterns that are exploited in the design of optimised video communication systems in this thesis.

Chapter 7 explores the motion characteristics and relative encoding requirements of BSL video image regions. This chapter explores the nature of the video image content which is given priority by the HVS at the point of gaze (identified in the EMT experiments) using optical flow estimation techniques and compares this with the priority (in terms of bit allocation) given to image content in a standard video coding scheme. This analysis is central to the development of the hypothesis that changing the priority given to coding different image content regions in BSL video, to meet the requirements and visual sensitivity of deaf viewers, reduces the bit rate of the BSL video images without loss of perceived quality for the user.

Chapter 8 describes the content-prioritisation algorithms developed in this thesis for BSL video coding. It describes the novel spatial and temporal video image foveation methods which pre-process BSL video image content based on the foveal response of the HVS (reviewed in Chapter 2) and the visual response of deaf people to BSL video communication (determined in Chapter 6).

The design and application of the methods developed in this thesis to measure the perceived quality of the outputs of the spatial and temporal video image foveation methods are given in Chapter 9. This Chapter identifies the maximum degree of spatial and temporal image foveation which provides acceptable quality for BSL communication.

The performance of the developed systems is compared with standard systems in terms of the perceived quality of BSL communication at fixed bit rates in Chapter 10.

Part Three contains the final discussion and conclusions of the thesis. Chapter 11 reviews the research problems and developments and suggests further work to develop research in this field. Chapter 12 summarises the conclusions for each of the research objectives and the original contributions of the thesis.

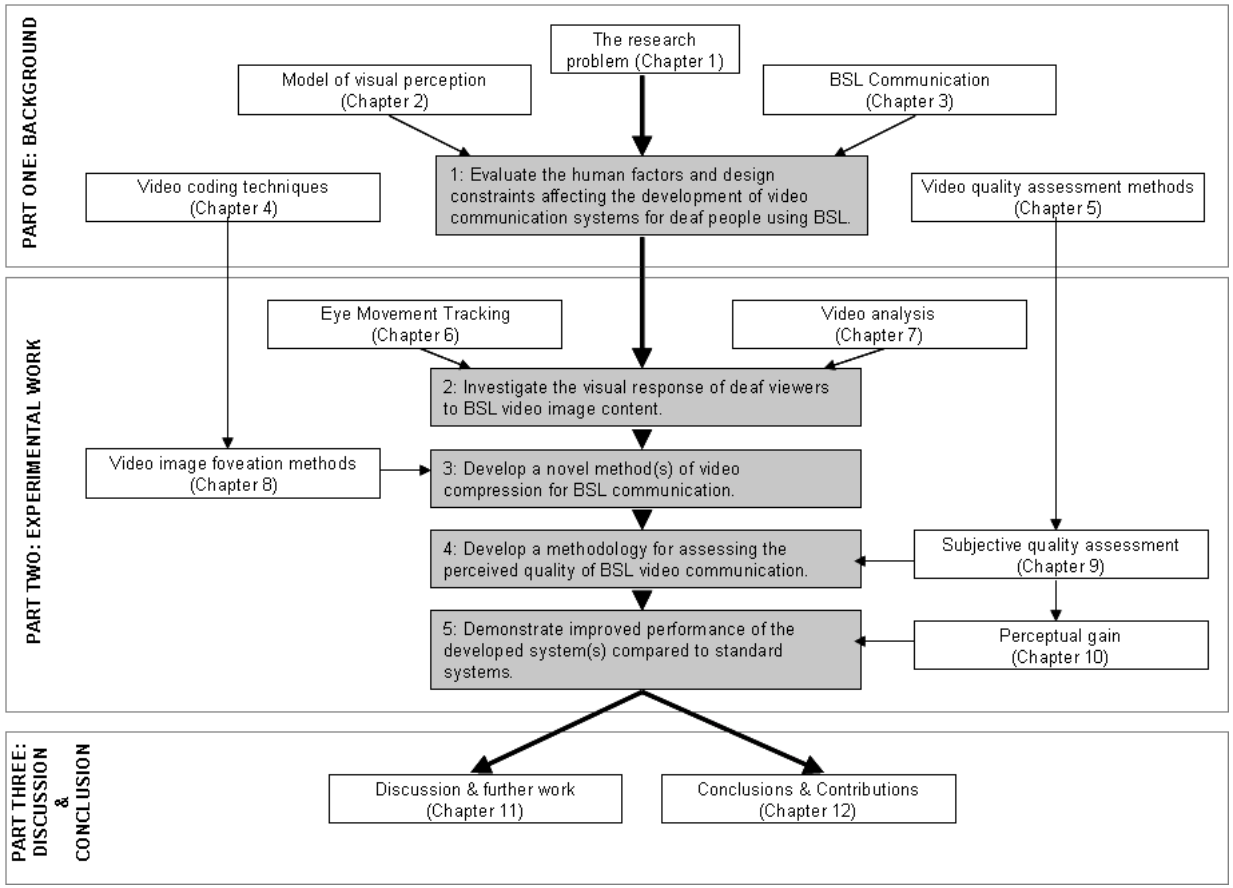


Figure 1.1 The Development of Content-Prioritised Video Coding for BSL Communication

Chapter 2: Human Vision

This chapter reviews current theory of the anatomy, visual signal processing and performance of the Human Visual System (HVS) and explores the visual attention mechanisms and the perception of visual information which are important in the design of video communication systems for deaf people. A model of vision is developed which identifies the important factors influencing the visual perception of BSL communication.

2.1 The Human Visual System

The Human Visual System (HVS) consists of the sensory apparatus and neural processing mechanisms which enable human vision.

"The eye is not simply a camera but a complex device for capturing and processing data. The aim of the data processing is not simply to exploit the data channels available in the HVS with greater efficiency but also to put the image data into a form suitable for interpretation by the brain." (Sharp and Philips, 1997)

The physical anatomy of the eye and the visual processing, performance and perception mechanisms of the HVS are explored.

2.1.1 Anatomy of the Eye

The eye is a spherical sense organ which channels the visual stimulus (Figure 2.1). Light enters the eye and passes through the ocular refractive media (Sharp and Philips, 1997). The ocular refractive media (cornea, lens and vitreous humor) are a collection of transparent structures. The light stimulus is refracted by the cornea, passes through the pupil and is focussed by the lens on the retina.

The outer wall of the eye contains three layers. The sclera is the tough fibrous protective layer of the eye. The choroid layer, containing the blood vessels, is modified to form the iris at the front of the eye. The iris controls the size of the aperture (pupil) of the eye and thus the amount of light reaching the retina (Simpkins and Williams, 1988).

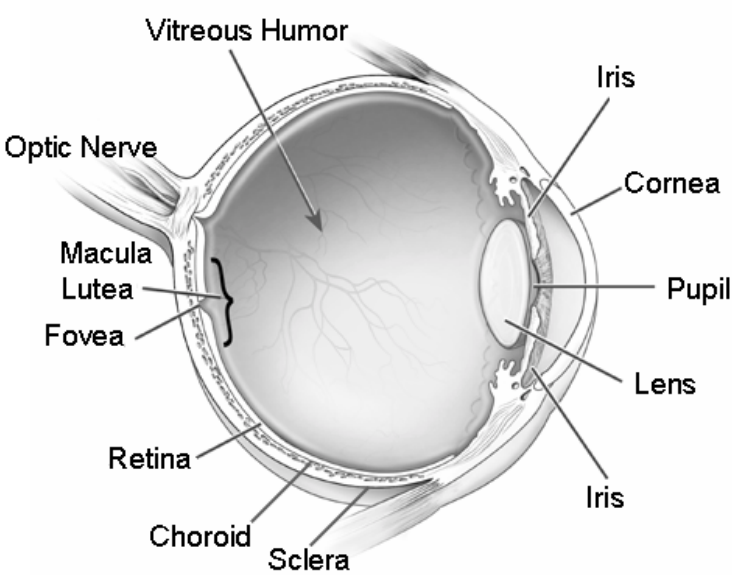


Figure 2.1.The Eye (Adapted from National Eye Institute, http://www.nei.nih.gov)

The retina is the inner layer involved in the detection of the visual stimulus. It contains approximately two hundred and fifty million photoreceptors that translate the external stimulus (light) into an internal neurosignal. There are two types of photoreceptor, rods and cones. Rods are distributed throughout the retina and are sensitive to low levels of stimulation and important for night vision. Cones are packed near the centre of the eye and need more intense levels of light. Cones are essential for colour vision and are most active in daylight vision (Gazzaniga, Ivry and Mangun, 1998). There are approximately 6.4 million cones and between 100 and 125 million rods in the human retina (Osterberg, 1935). The retinal surface is generally flat but contains a shallow pit of diameter 1500 micrometers called the fovea corresponding to a visual angle of 5 degrees (Findlay and Gilchrist, 2003).

The fovea (Figure 2.1) is a specialised region of the retina located in the posterior of the eye in the centre of the *macula lutea*[2] (Simpkins and Williams, 1988). It has a

[2] *Macula lutea* - a yellow spot at the centre of the retina.

particularly important role to play in visual perception since vision is not equally acute in all parts of the visual field. The fovea enables acute vision (referred to as macular or foveal vision) since the photoreceptors (cones) are at their highest density in this region. The density of cones at the centre of the fovea (50×50 μm) is 147,000/mm^2 and there are no rods in the central 200 μm (foveal area) of the retina (Osterberg, 1935).

Visual cells (apart from the cones) are displaced towards the periphery of the retina, away from the foveal pit. This results in a thinning of the retina in the foveal region that improves the optical quality of the image on the photoreceptors (Findlay and Gilchrist, 2003). In the very centre of the foveal pit is a region known as the foveola. Cone density continues to increase to the very centre of the foveola so that inter-cone spacing is approximately 2.5 micrometers, decreasing to 5 micrometers at 1 degree eccentricity (Hirsch and Curcio, 1989). Eye movements (discussed further in Section 2.2) direct the fovea to the part of the visual scene requiring high image resolution. Visual acuity afforded by the peripheral region of the retina, outside the fovea, falls rapidly and smoothly away from the centre of vision. Resolution is reduced by a factor of 2 at 2.5 degrees from the fixation point and by a factor of 10 at 20 degrees from the fixation point (Hendee and Wells, 1997). The visual effect of foveal vision is illustrated in Figure 2.4 and discussed in Section 2.3.2 in the context of the spatial response of the HVS. Peripheral vision is specialised for night vision and for processing gross movement in the visual scene where high resolution of the image is not required. Studies of deaf people at the University of Oregon (Gazzaniga, Ivry and Mangun, 1998) using ERP (Event-Related Potentials) and fMRI (functional Magnetic Resonance Imaging) revealed that deaf subjects are faster and more accurate at detecting moving targets in the peripheral visual fields than hearing subjects. This was evidenced by enhanced visual processing in the temporal lobe of the brain; that is, enhanced processing ability rather than an enhancement of the visual apparatus. The explanation for this enhanced ability is that deaf subjects rely on peripheral vision for processing gestures away from the point of gaze during a multi-channel sign language conversation (Siple, 1978), discussed in Chapter 3.

The limitations of the visual system in terms of visual acuity are an important consideration in the design of visual information systems. In addition to optical aberrations and differences in optical resolution, there are physical limitations related to

eye movements (for example, rapid eye movements and unstable fixations explored in Section 2.2) and a defect of the retina known as the ‘blind spot’. The ‘blind spot’ is a region of the retina that contains no photosensitive cells and so no vision is facilitated in this region. It subtends a visual angle of 3-5 degrees and is the point in the optic disk (Figure 2.1) at which the axons and blood vessels come together and leave the ocular globe (Simpkins and Williams, 1988).

Visual acuity (the ability of the observer to see high contrast spatial information in an image) may be limited by optical blur due to physiological factors/problems in the HVS. Acuity is tested using the Snellen Visual Acuity chart (measure of optical quality) and corrected with optical lenses. However these traditional ‘eye tests’ are designed to measure efficiency of focal vision only. The Snellen letters measure ability to resolve and recognise small shapes (optotypes) in the centre of vision. While this is a necessary measure of visual ability, it provides no information about other relevant visual processes. Measures of the sensory qualities of vision (fine acuity, perimetry, and the ability to discriminate colours) are useful but the most important contributors to the design of systems are likely to be perceptual, involving more complex and sophisticated ways of extracting and interpreting information. Snellen charts sample only part of the visual system's capacity. Contrast sensitivity and the spatio-temporal response of the HVS are the main visual characteristics which are important in the design of systems which operate within visual perception limits (discussed in Section 2.3).

Despite the limitations of the HVS, normal (or corrected-to-normal) vision appears smooth and uninterrupted. Visual processing strategies and perception mechanisms ensure that humans are not aware of physical visual limits.

2.1.2 Visual Processing

The mechanisms for visual information processing in the eye reveal strategies for efficient processing within physical limitations. The HVS does not create an ‘internal screen’ or ‘scale model’ representation of our visual environment in the brain (O’Regan, 1992; Sharp and Philips, 1997). Compensatory mechanisms and extensive signal processing of visual information occur in the eye and brain in a way that precludes the necessity for an internal photographic representation of the outside world in the HVS

and minimises the amount of information which must be processed in the nervous system.

The neural pathways from the eye to the brain, which facilitate processing of a visual stimulus, are illustrated in Figure 2.2.

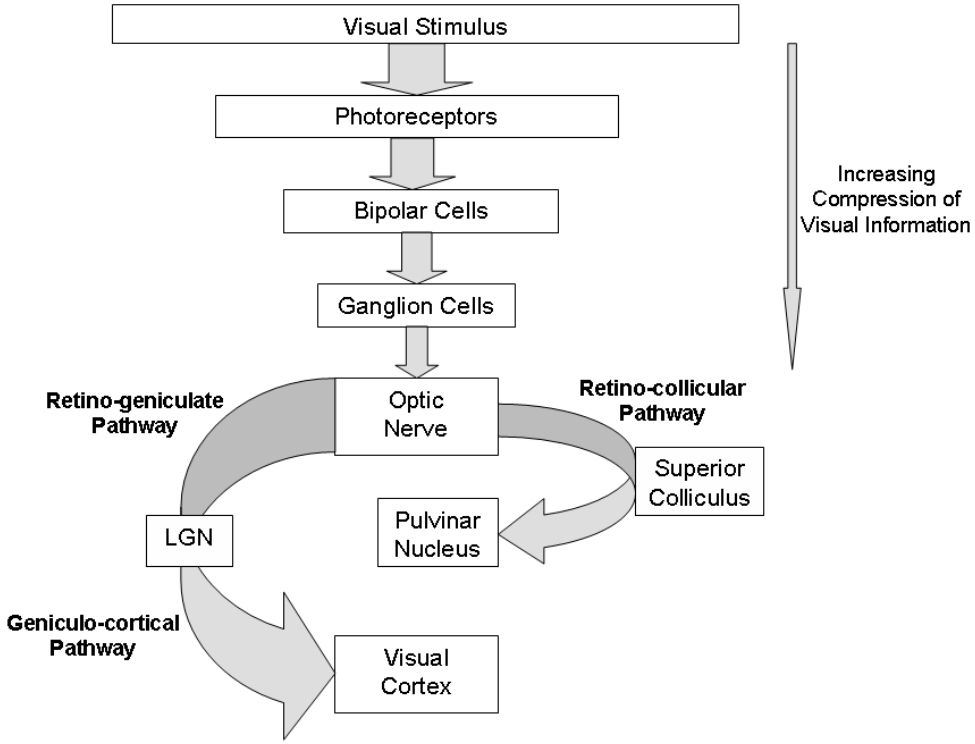


Figure 2.2 Visual Processing

Photoreceptors initiate neural processing of the visual signal. The output from the photoreceptors is transmitted to bipolar cells and then to approximately two million ganglion cells (Simpkins and Williams, 1988) and so there is extensive convergence or *compression* of visual information by a series of low-pass and high-pass spatial and temporal filters (discussed in Section 2.3.1). Axons of ganglion cells meet to form a bundle in the optic nerve (Figure 2.1) which transmits visual information to the Central Nervous System (CNS). Visual information is transmitted in the form of neural signals from the eye to the visual cortex of the brain by three visual pathways (Gazzaniga, Ivry and Mangun, 1998). The retino-geniculate pathway, from the retina to the Lateral Geniculate Nuclei (LGN) in the thalamus, contains more than ninety percent of the axons in the optic nerve. The LGN provides input to the visual cortex via the geniculo-

cortical pathway. The retino-collicullar pathway comprises the remaining ten percent of optic fibres. This pathway innervates the superior colliculus (in the midbrain) and pulvinar nucleus (in the thalamus) and plays an important role in visual attention, discussed further in Section 2.2.4.

High-level visual centres in the brain are required to recover (decompress/decode) the detail of the visual scene from the visual information signals. The visual information that the transmitted neural signals represent is distributed across different subsystems for processing in the brain. There are approximately thirty-two prominent visual areas of the brain. Processing through these areas is not sequential; feedback occurs in the system. Visual perception is an analytical process since each visual area provides a map of external space. The maps differ with regard to the information they represent, for example colour or motion. Neurones are responsive to change (contrast), for example cortical neurones are responsive to edges. They code information about attributes and location of objects. These cells are specialised even to the level of speed and direction of motion. Cells highly sensitive to motion are also highly sensitive to contrast (brightness) and cells highly sensitive to colour are not highly sensitive to other features; for example, contrast, location, motion, orientation and depth (Gazzaniga, Ivry and Mangun, 1998). Studies of the visual perception of motion, luminance and colour by Morland, Jones, Finlay, Deyzac, Le and Kemp (1999) found that cortically blind subjects were able to respond to colour and motion (but not brightness) features of visual stimuli; a phenomenon called ‘blind sight’. This is further evidence that the perception of different stimulus attributes is a function of different visual areas of the brain and that overall visual perception is analytical rather than an exact internal representation of the stimulus.

Visual systems are characterised by multiple pathways, each with individual specialisations. Information is processed and analysis conducted concurrently (Gazzaniga, Ivry and Mangun, 1998). Neuro-physiological evidence for concurrent processing using PET (Positron Emission Tomography) enables the link between behaviour and anatomy to be explored by identifying brain regions active during tasks. Gazzaniga, Ivry and Mangun (1998) cited an example of a study that showed different visual areas were activated when processing motion and colour information. This concurrent processing ability was evident during a visual search. Visual search

requires analytical processing of image features relevant to the task (for example searching for an 'L' shape amongst an array of 'T' shapes, by shape/orientation detection). There is no conscious search; the targets 'pop out'. 'Popout' is the ability to recognise the unique stimulus property (for example shape or orientation) of the target object in familiar tasks (Gazzaniga, Ivry and Mangun, 1998). It was postulated that, as BSL communication is a familiar task for experienced sign language users, deaf people will have the ability to recognise the unique shapes of BSL gestures formed by the hands and fingers.

Given the sensitivity and responsiveness of the HVS, it is evident that it must perform a high degree of image processing to minimise the amount of information that must be processed. The HVS uses strategies to compress the visual data to signal changes in the information rather than transmit all the information about the visual scene. The 'on-centre', 'off-surround' types of receptors in the HVS effectively act as a band pass filter, "removing dc and low spatial frequency information from the scene and transmitting information about the change in the spatial data in the retinal image" (Sharp and Philips, 1997). Visual filters are described in Section 2.3 in the context of the spatio-temporal response of the human eye. In addition, specialist cells that respond to specific features in the image (such as lines at a particular orientation) reduce the amount of information that has to be transmitted. The mechanisms for recognition and perception of image objects and features are described in the following section.

2.1.3 Perception of Objects and Features

Perception of image features and recognition of specific visual objects (for example, the face and hands in BSL communication) are important aspects of visual processing as they enable more efficient decoding of visual information based on knowledge and experience in relation to the visual task.

It is generally accepted that examination of a visual stimulus involves two phases (Gale, 1997). The first phase is a pre-attentive glimpse. Pre-attentive processing is automatic and employs parallel processing to get a global impression of the scene and leads to detection of physical features anywhere over the stimulus, previously described as 'popout' (Section 2.1.2). This is followed by a second, focal attention, phase consisting of an active serial visual search using eye movements. In this phase information is

processed from the foveal region and an area around it known as the 'functional visual field'. The density of irrelevant display items affects the size of this area as do target and non-target similarity for simple stimuli. The functional visual field shrinks as foveal load increases but experiments by Ikeda & Takeuchi (1975) demonstrated that experience and training can overcome this shrinkage. This may be the case for experienced sign language users; that is, deaf people may be able to filter out irrelevant detail and maintain a wider functional visual field during sign language communication. The analysis of visual information by image features or holistic systems, for object recognition, depends on the visual task. This is important for a specific skilled task such as BSL communication.

2.1.3.1 Recognition of Image Features

Feature detection is facilitated by specific nerves that fire only when excited by a particular feature (for example; colour, form, vertical line, angled line, circular blob, movement in a particular direction). These feature detectors enable the selection of specific features from the flood of information detected by the eye (Kelsey, 1997).

Some simple features or 'textons' (Julesz, 1986), for example; colour, line orientation, line intersections, are detected during the pre-attentive phase. Texton differences permit the target to be recognised and give rise to head/eye movements to bring the target on the fovea in the next, focal attention, phase of vision (Kelsey, 1997). In this phase, foveal vision is used to scan with attentive search. Fiorentini (1989) investigated whether visual processing in foveal vision was different than in non-foveal vision. Her experiments revealed that the time for search was the same for foveal and non-foveal vision and that a specific task (counting objects in this case) required an attentive search, not just texton detection. The importance of attention during execution of a visual task was highly relevant to this thesis as a deaf viewer is performing a very specific, specialised linguistic task. Visual attention is explored in more depth in Section 2.2.4.

Studies of the damaged brain have facilitated discovery of the processes required for perceiving and recognising image features. The strongest segregation of function is motion and colour perception. Gazzaniga, Ivry and Mangun (1998) described a stroke victim (PT), with good vision, who was unable to visually recognise his wife until she

spoke. He recognised his wife's movement and voice but not colour or form information about her. The same subject was unable to recognise a character (colourful impressionist style) in a Monet painting as a person but was able to recognise a woman in a cubist-style painting (which is more abstract but simple in its hue and form). Colour helps to define regions of visual space; a patient with no colour vision can detect brightness but not hue. Patients with deficits in motion perception (*Akinetopsia*), view the world in snapshots. Gazzaniga, Ivry and Mangun (1998) described a patient (MP) with colour and form perception intact but with impaired judgement of direction and speed. This patient could not detect motion at speeds greater than 20 degrees/sec and saw change information only. The motion pathway is colour-blind and colour information is processed in a different region of the brain. Feature recognition is therefore a specialist function of designated brain regions.

2.1.3.2 Perception and Recognition of Visual Objects and Faces

Humans look directly at objects they wish to identify, taking advantage of the greater acuity of foveal vision. Recognition of specific visual objects is a higher perceptual function in the HVS. The specific responsiveness of individual cells to complex stimuli (for example, hands, faces) is facilitated by 'grandmother' higher order neurones and a holistic view is obtained by what is known as the ensemble theory derived from a study of visual disorders (Gazzaniga, Ivry and Mangun, 1998). *Visual Agnosia* is the inability to recognise objects in a subject with otherwise good visual acuity (for example, after a stroke). In contrast, patients with *Optic Ataxia* recognise objects but can not use the visual system to guide their actions. In this case eye movements are not controlled by spatial knowledge, resulting in inappropriately directed saccades (rapid eye movements) so that the target object is not in the foveal area. This disorder is a result of lesions of the parietal cortex. The parietal lobe plays a critical role in selective visual attention (explored further in Section 2.2.4) and so the link between attention and object recognition is important. *Apperceptive Agnosia* is the inability to recognise objects due to computational problems (for example, following carbon monoxide poisoning) whereas *Prosopagnosia* is a deficit in the ability to recognise faces that cannot be directly attributed to deterioration in the intellectual function. This deficit is specific to the visual modality (like *Agnosia*); the patient will be able to recognise the person from the voice but will not recognise the person's face. These conditions

demonstrate that inability to recognise faces is not always related to object perception, suggesting a different processing mechanism for faces.

Face perception depends on a specialised processing system involving temporal lobe activity. This results in a holistic representation of the face; the component parts are not analysed (and so the observer may not notice that the moustache on a recognised face has been shaved off). This phenomenon is known as 'inattentional blindness' (Mack and Rock, 1998; Bruce, Green and Georgeson, 2003). An incongruous or unexpected object appearing in a display may not be noticed by observers if their attention is diverted elsewhere by a task. Attention to different parts of the scene to carry out a task determines whether observers report seeing or being aware of anything at all.

In contrast to basic object recognition, face recognition (individual identity) is not successful if based on simple edge features alone; it requires information about surface characteristics (such as pigmentation) and texture (such as skin and hair) of facial features (Bruce, Green and Georgeson, 2003). Small facial details, such as eyebrow movements, give important clues to the meaning of verbal communications as well as sign language (visual) communications. In addition to facial posture, the timing of face movement is important. Some expressions flicker rapidly across the face and last as little as 200 milliseconds. Timing affects the accuracy of expression perception and the perceived intensity of the emotion. Altering the temporal properties of the display affects the emotional impact of the message (Bruce, Green and Georgeson, 2003).

Evidence that deaf people have superior face discrimination ability due to reliance on facial expression for linguistic contrasts (discussed in Section 3.1.3) also demonstrates the importance of mechanisms for face recognition. In addition, deaf-from-birth subjects have been found to process visual stimuli in the auditory cortex of the brain (Finney, Fine and Dobkins, 2001). This neural reorganisation of the visual sensory modalities was determined using fMRI and is related to the processing of motion in the brain. This further enhancement of the HVS in deaf people, to facilitate processing of information important for visual communication, is an indicator that a system designed specifically for the visual requirements of deaf signers is appropriate.

Visual perception is based on analytical processing of visual data, received from the eye, by the brain. What is 'seen' is based on visual information and familiarity with, or

experience (a knowledge-base) of, the visual task. Sign language communication is an active visual and cognitive task requiring interpretation of gestures and knowledge of language structure and production. Studies of feature and object recognition, without consideration of the cognitive task and attention mechanisms for active vision, would have been very limiting for this thesis and so a more detailed examination of eye movements and attention mechanisms for active vision is explored.

2.2 Eye Movements and Attention

Humans are not passive processors of information. “We can select from the dazzling array that impinges on the senses at any time. Depending on our goals, the relative importance of different sources of information constantly changes.” (Gazzaniga, Ivry, and Mangun, 1998)

A detailed investigation of the viewing behaviour of deaf people necessitated the study of eye movements and visual attention during video scene perception. This section of the thesis reviews eye movements, factors affecting eye movements, fixations, scan-path theory, active vision and attention.

2.2.1 Eye Movements

The fovea is the region of the eye which is responsible for detailed processing of visual information (Section 2.1.1). Head and eye movements direct this area of high visual acuity to sample the visual stimulus (Gale, 1997). Foveal information is clear and fully chromatic whereas peripheral information is blurry and weak in colour to a degree depending on the distance from the fovea. In order to obtain high resolution information about the spatial and/or chromatic attributes, the visual scene must be explored using eye movements to place different information in the fovea at different times.

Oculomotor control systems, including involuntary eye movements (Physiological Nystagmas) and the four major classifications of voluntary eye movement (saccadic, smooth pursuit, vergence and vestibular), were reviewed by Robinson (1968) in the context of application to engineering problems (including gunfire control and radar tracking). Study of these eye movement types is also important for understanding how deaf viewers sample the visual environment.

Physiological Nystagmas are constant, small amplitude, non-selective eye movements (Palmer, 2002). When the eye fixates on a stationary object, three types of small amplitude, involuntary eye movements (slow drift, fast flicker and a superimposed high frequency tremor) occur. These continuous movements (drift and flicker) and superimposed tremors (up to 150 cycles/second with amplitude of up to approximately half cone diameter) ensure that the retinal image always changes even if the scene does not and so there is no perceived fading of the image during a fixation (Sharp and Philips,1997).

Saccades are the most frequently executed type of voluntary human eye movement. Several billions are made in a lifetime (Gale, 1997). They are rapid eye movements which are executed to visually scan the scene and fix a new object of interest on the fovea (Cumming, 1978). Saccades are conjugate[3], ballistic in nature, may be curved or linear and are normally less than 15 degrees in amplitude (Gale, 1997). According to Palmer (2002), a saccade takes approximately 150-200 milliseconds to plan and execute and reaches an angular velocity of up to 900 degrees/second. Gale (1997) reports saccades require 250 milliseconds to plan, approximately 50 milliseconds to execute, reaching very high velocities of up to 600 degrees/second. Saccades frequently over or under-shoot the target triggering corrective saccades. Saccades are unbiased real-time estimators of predicted motion, evidenced in a study of real-time detection of unseen arm reaching movements (Ariff, Donchin, Nanayakkara, and Shadmehr, 2002). There is no salient disruption or visual perception of blurring during the rapid shift of gaze produced by a saccade. This is due to a phenomenon known as 'saccadic suppression' (Palmer, 2002). During a saccade, there is suppression of uptake of visual information (Gale, 1997). During this time information is processed for trans-saccadic processing (Findlay and Gilchrist, 2003). Saccadic eye movements are the most important type in the study of information processing tasks (Rayner, 1998).

Smooth pursuit eye movements are used to track the position of a moving target to keep it in foveal (acute) vision. This type of eye movement cannot be made in the absence of a moving visual target. It is slower than a saccade (30-100

[3] Conjugate eye movements are coordinated motions of both eyes (Sharp and Philips, 1997).

degrees/second), smooth and continuous (not ballistic) and corrected based on feedback. Visual acuity is high for the image of the tracked object. Untracked and stationary objects appear to be blurred due to relative motion on the retina. There is a trade-off between accurate tracking and acuity (Palmer, 2002).

Related to pursuit movements are compensatory movements which adjust for head and body movements. These can compensate for a movement rate of approximately 1 to 30 degrees per second.

Vergence movements are disconjugate eye movements which allow both eyes to focus on the same target. The eyes converge to look at nearby objects and diverge to focus on distant ones. Vergence movements are much slower than saccades (approximately 100 degrees/second) and last approximately one second (Gale, 1997).

Vestibular eye movements occur when the eyes rotate to compensate for head and body movements in order to maintain the same direction of vision (Rayner, 1998).

2.2.2 Fixations

The HVS uses conjugate eye movements to fixate an eccentrically located object of interest on the fovea (using saccades) and to follow a moving object. This is achieved by matching its velocity so that foveal fixation is maintained, using pursuit movements (Sharp and Philips, 1997). Viewing a scene involves a sequence of saccades and fixations. Fixations occur between saccades, during which the eye dwells on an object for a variable period of time. The average duration of a fixation is 300 milliseconds (Palmer, 2002). The main variables of a fixation are duration and location.

During a fixation, the viewer encodes information about the visual stimulus from the current eye position and programs the subsequent saccade. For example, in a reading task, approximately the first 50 milliseconds of fixation time is spent encoding information, therefore the majority of time is for preparation of the next movement. Experienced searchers, for example text readers (and by implication sign language users) make fewer fixations of longer duration than observers performing simple untrained tasks (Gale, 1997).

Factors affecting the duration of fixations during picture viewing have been investigated by vision researchers. Loftus (1985) showed that presentation of a line drawing at low contrast produced an increase in mean fixation duration compared with a normal contrast view. Low-pass filtering (an operation which removes high spatial frequency, fine detail information from an image producing a blurring effect) was shown to produce a similar effect on fixation time (Mannan, Ruddock and Wooding, 1995).

The duration of a fixation is also considered in terms of gaze or dwell time on a particular location. Perceiving a realistic visual scene generally requires a sequence of many different fixations (Findlay and Gilchrist, 2003). The gaze period is composed of individual fixations grouped together within a spatial threshold value. Fixations occur on areas of high information content (for example, contours rather than homogeneous areas) and/or regions of interest to the viewer (Gale, 1997). Guo, Mahmoodi, Robertson and Young (2006) found that viewers spent more time looking at faces in images. Fixation times on implausible objects in the scene context have also been shown to be longer (Findlay and Gilchrist, 2003).

2.2.3 Fixation Patterns and Scan Path Theory

The sequence of eye movements and fixations plays an important part in the perception of visual stimuli.

Early vision scientists used crude eye tracking methods to study the viewing behaviour of subjects presented with complex still images. These methods involved the application of photographic techniques to record eye movements during reading and picture viewing. Buswell (1935) reported that a scan was used to form a general impression of the scene. Yarbus (1967) used a suction cap fitted to the eyeball with a small mirror attached so that a beam of light, reflected from the mirror, moved with the eye and was captured using photographic techniques. He found that additional viewing time spent on a still image was not used to examine the whole scene but was used to re-examine the most ‘important’ elements of it. By superimposing eye movements on the stimulus picture, he determined which parts of the image observers found most informative. He observed different scan paths were obtained when users were given specific instructions. He concluded, like Buswell (1935), that the way in which the eyes

explored a complex image depended on the nature of the task. These researchers identified that the cognitive aspect of the visual task affected viewing behaviour.

Noton and Stark (1971) developed a hypothesis which proposed that a new visual stimulus produces a particular fixation sequence (or scan path) and that this recurs when the same stimulus is recognised when re-presented to the viewer[4]. Scan path theory was refined by the work of Groner between 1984 and 1988 (Gale, 1997) who defined local and global scan paths. Local scan paths were described as those which "reflect the spatio-temporal organisation of the fixations on a local scale of successive events" and global scan paths as those which "reflect the distribution of eye fixations when taking into account the entire inspection process". It was also found that these scan paths were subject-specific and that global scan paths in particular were related to the viewer task.

Eye Movement Tracking (EMT) equipment, employing a range of techniques, has allowed researchers to determine the specific sequence of fixations that observers execute when exploring a visual scene. No single technique is appropriate for every situation and the choice depends on the task, resolution, accuracy required and sampling rate of the technique. EMT techniques, methodological issues and the application of EMT in the experimental work in this thesis are detailed in Chapter 6.

Vivianni (1990) dismissed the use of eye movement data obtained from EMT unless it was studied in parallel with an articulated theory of cognitive activity for the task in question and argued that only then could it provide useful information about visual perception. He postulated that scene information is available to the viewer at a number of different levels or 'spatial scales' and that eye gaze can not measure whether a particular 'scale' is selected. He stated that 'what meets the fovea is only part of what reaches the mind's eye', implying that the direction of eye gaze does not necessarily correspond with the focus of attention. The importance of task to viewing behaviour, highlighted by Vivianni (1990), was relevant to the study of eye movements for the complex visual task of BSL communication and influenced the design of the EMT

[4] This phenomenon was tested in the first set of EMT experiments applied in this research (Section 6.3) and found to be true for deaf people viewing sign language video.

experiments and the analysis of the data obtained from deaf viewers of BSL video (Chapter 6). In addition to the visual task, other factors affecting eye movements and fixations also need to be considered when analysing eye movement data. These include real world knowledge (physical laws, experience and expectancies), emotional and social factors, and the instructions and training given to the subject. These factors were also considered in the design of the experimental work and the analysis of data from eye tracking and subjective quality testing.

There are many examples of the successful use of eye tracking techniques to investigate the visual processing of information in cognitive tasks. Land, Mennie, and Rusted (1999) used eye tracking to investigate eye movements during active tasks, including driving, table tennis, piano-playing and tea-making. Their results demonstrated that gaze is directed to the points of the scene where most information can be extracted and that the eye anticipates movement rather than follows it. Hyőnä and Lorch (2004) used eye movement tracking to study the effect of topic headings on text processing in reading tasks and found that these acted as powerful signalling devices in the processing of information. A comprehensive review of eye movement studies of reading and other information processing tasks (including typing, visual search, scene and face perception) by Rayner (1998) revealed different mechanisms depending on the task. Despite the prevalence of moving image applications, this review identified that there has been very little research on gaze patterns in this field. Guba, Wolf, DeGroot, Knemeyer, van Atta, and Light (1964) found that television viewers had a strong tendency to fixate on the face of a narrator even in the presence of a distracter. However, most of the scene perception research reported in the literature before 2002 (when the research for this thesis was initiated) concentrated on viewing behaviour for still image stimuli. Other applications of eye movement tracking research include user interface control and gaze-based interfaces (Jacob; 1991, 1993, 1994, 1995) and for user interface evaluation of computer applications and web sites (Jacob and Karn, 2003)

In addition to the study of foveal vision by EMT, research on object perception performance in peripheral vision has revealed some important information about how image content is perceived in low resolution vision. Shiori and Ikeda (1989) measured object perception performance in peripheral vision using a combination of gaze

contingent windows and filtering techniques. They found that even very low resolution detail, well below the acuity limit, in the peripheral location aided performance considerably. Van Diepen, Wampers and d'Ydewalle (1998) developed techniques involving a moving window and Fourier filtering and found that performance was better when the peripheral material was high-pass filtered (removal of low spatial frequency, gross detail information) than when it was low-pass filtered. This was thought to be a surprising result at the time but it demonstrated the potential benefits of pre-filtering image content to improve peripheral visual performance and the need for further research to identify the visually important characteristics of video image content.

2.2.4 Attention and Selective Visual Perception

Voluntary eye movements are the main instruments of selective attention. Attention is global (to the whole scene), to a selected object or set of objects, to a specific part of an object or to the property of an object (for example, colour). Mack and Rock (1998) proposed that attention is required for conscious perception of anything at all. A rich model of vision assumes that active humans engaged in active vision and requires the study of attention mechanisms (Findlay and Gilchrist, 2003). Attention involves more than sensation and perception, it can be directed to internal mental processes (for example, memory or a complex task). Gazzaniga, Ivry, and Mangun (1998) referred to the definition of attention given by the psychologist William James in 1890: "the taking possession of the mind, in clear and vivid form, of one out of what seem several simultaneous possible objects or trains of thought". Attention is selective given the array of possible targets and humans can direct attention, covertly or overtly. How these attention mechanisms are applied depends on the model of vision, passive or active.

Traditional vision science theory is based on a passive model of vision which does not take account of eye movements. Early experiments, reviewed by Findlay and Gilchrist (2003), were designed to determine visual thresholds and were conducted using brief displays during which eye movements were prevented. This approach assumed that the internal representation of the scene is a processed representation of the retinal image and did not account for the non-homogeneity of the retina. It also led to assumptions about the way visual attention is conceived. Covert attention, the ability to

attend to part of the visual array without moving the eyes, has been the main focus of research on visual attention mechanisms. A passive model of vision considers covert attention to be a 'mental spotlight' that can be directed to any part of the 'internal image' and that covert attention is the main means of attentional selection.

An active model of vision recognises the contribution made by gaze shifts to visual perception and cognition. It also takes account of the non-homogeneity of the retina, in particular the fovea as a region of high visual acuity and the point at which vision is centred. Overt attention, the ability to saccade and foveate part of the visual scene, is the basis of 'active vision' research (Findlay and Gilchrist, 2003). Active vision researchers argue that spatial selection of visual information by fixation gives greater processing advantage. In addition, it is possible to shift attention without moving the eyes to attend to different peripheral objects while maintaining eye fixation (Posner, 1980). Accurate measurement of where an observer is looking is not always a measure of attention (Shepherd, Findlay and Hockey, 1986). It is possible to make an eye movement without producing a change in attention locus. Gale (1997) reported that eye movements are not necessary for a change in attention for some tasks (for example, touch typing). This covert attention confers minimal processing advantage and is thought to be closely related to overt saccadic selection. Covert attention allows eye movements to be guided to relevant aspects of the scene for the task.

The relationship between eye gaze and attention is a complex one and there are conflicting views on the strength of their association. The location of attention can not be observed in the same way that gaze points can and therefore must be inferred from indirect evidence (Bruce, Green and Georgeson, 2003). Bruce *et al* (2003) argued that, although in some circumstances spatial attention may be allocated to a different part of the visual field from that fixed in foveal vision, attention and the control of gaze are closely interrelated. Saccades are preceded by shifts of attention to the region of the next fixation guided by covert attention in the peripheral field of vision (Groner and Groner, 1989).

Rayner (1998) presented evidence that, for complex stimuli, it is more efficient to move the eyes than to move attention and he supported the argument that attention precedes a saccade to a given location in space. Rayner (1998) also argued that locus of attention and eye location could be decoupled in simple discrimination tasks but in

complex information processing tasks, such as reading, the link between the two was very strong. Hoffman and Subramaniam (1995) presented further evidence for this in studies which found that subjects were not able to move their eyes to one location and attend to a different one in target detection. At the neural level, functional Magnetic Resonance Imaging (fMRI) studies by Corbetta, Akbudak, Conturo, Snyder, Ollinger, Drury, Linenweber, Petersen, Raichle, van Essen and Shulman (1998) provided evidence that attentional and oculomotor processes are tightly integrated. The human response to a visual stimulus depends on many factors but is ultimately task-specific (Findlay and Gilchrist, 2003; Gale, 1997), implying that how we see depends on the task being performed. This theory was supported by Henderson and Hollingworth (1999) who stated that complex interrelated factors influence high level scene perception. They identified three main factors:

1. The influence of cognitive processing on the position and duration of a fixation during a task. Perception is directly related to the visual task. This supports the central theory for this thesis, that there is no single 'generic' mode of visual communication and adds weight to the argument for a system designed for the specific visual task requirements of the viewer.
2. The nature of scene representation during saccades and other brief time intervals.
3. The relationship between scene and object perception, particularly the impact of scene semantics on the identification of objects within it. In addition, Chua, Boland and Nisbett (2005) found cultural (East/West) variation in viewing patterns for natural scene perception.

The impact of context, task, eye movements and attention (including conflicting theories regarding the strength of the relationship between eye location and focus of attention) has important consequences in the context of this research. The location of eye movements and attention mechanisms are important factors in the optimisation of video communication systems for BSL users. It was assumed that, as BSL is a complex information processing task (explored in Section 3.1), eye gaze and locus of attention are the same.

2.3 The Spatio-Temporal Response

Girod (1992) identified spatio-temporal properties of the HVS as being of fundamental importance to the design and evaluation of image communication systems. The spatial and temporal response of the human eye and the filtering properties of the HVS are important factors in the design of video communication systems optimised for visual perception mechanisms.

2.3.1 Spatial and Temporal Filtering

The HVS facilitates reduction of data at all stages of processing. Visual information is produced by selective processing and suppression (filtering) of certain types of visual data (Ginsburg and Hendee, 1997).

Different types of receptor cells in the visual cortex respond to structures of different sizes (spatial frequency), contrasts and orientations. Some cells respond to large size objects of low spatial frequency, others respond to small objects of high spatial frequency. Some respond to high contrast and others to low contrast. Cells 'tuned' to low spatial frequency stimuli provide visual information about the gross features of an object. Cells tuned to high spatial frequency features provide information about fine details of an object. Thus, when stimulated by the same object, some cells provide information on gross features and others on fine detail of the object.

The retina acts as a temporal and spatial filter of patterns of light intensity and spectral compositions (Bruce, Green and Georgeson, 2003). The input and output of a filter are related by a transfer function which specifies how effectively different frequencies pass through the filter. In high-pass and low-pass filters only the temporal or spatial frequencies above or below a threshold value are transmitted, while in a band-pass filter only those frequencies within a particular band of values are transmitted. A high-pass filter suppresses low frequencies, giving sharper edges. A low-pass filter produces blurring. The retina is just one of a sequence of filters operating on the optical array.

The sequence of filters in the HVS is described by Bruce, Green and Georgeson (2003). Before photoreceptor signals pass to the neural cells of the retina, the

processes for image formation and transduction of light have already filtered out the high spatial and temporal frequencies. Optical formation of a retinal image and the pooling of light across each receptor aperture both act as low-pass spatial filters. The probabilistic nature of light capture implies a low-pass temporal filter which can be significant at low light intensities. These early filters are then followed by the neural processes of the retina which, in general, act as high-pass spatial and temporal filters through a process of adaptation (the main means by which humans adjust the sensitivity of photoreceptors and neurons in the retina to varying sensitivities of light) and lateral inhibition (centre surround antagonism, that is inhibition of light stimulation of surrounding cells) respectively. The spatial summation of light responses within receptive field centres is another low pass operation similar to the blurring imposed by the optical filtering in the eye.

There are multiple spatial frequency-tuned filters in the HVS that deal with important information in the image which can exist at any scale from course to fine. Rendering a picture in blocks makes recognition of features/objects difficult or impossible (used by TV producers to mask identities of faces in ‘sensitive’ footage). Recognition can be improved by blurring the image or viewing it from a greater distance. These actions filter out the high spatial frequencies that form the block edges and reveal the useful information in the lower spatial frequency information, similar to a blurred version of the original image. The reason that useful information can not be obtained from the blocked image is that conscious perception does not have access to the individual filters at different scales, only to the feature representations produced from them at a higher level of visual processing. The importance of the multiple scales (sizes) of filters is recognised in vision research but there is no single model for how the information is used. At small scales (high spatial frequency) the filters are more tightly tuned for spatial frequency (and orientation) than any of the proposed vision models would require and so there are questions which remain about their role. One theory, applied in coding schemes in digital applications for efficient coding, storing and transmission of image data (for example, in JPEG and MPEG), is that the whole set of filters, tuned jointly to different orientations and spatial scales, delivers a rather general but compact coding of local patches of the image (Bruce, Green and Georgeson, 2003).

The filtering characteristics of the HVS can be partly explained in terms of contrast sensitivity (the reciprocal of threshold contrast). Contrast sensitivity is a measure of the degree to which the visual system can discriminate between adjacent areas of light and dark, and usually includes a dimension of size (spatial frequency). Unlike measures of static focal acuity such as the Snellen test (Section 2.1.1), measurement of contrast sensitivity does not yield a single figure describing performance. The human visual system is discretely sensitive to information of different sizes from the optic array. Size in the visual sense is most commonly expressed in spatial frequency (cycles per degree). Low spatial frequencies provide general position, medium spatial frequencies provide general shape, and high spatial frequencies provide edges and fine details. The HVS functions in a similar manner to a crude Fourier Analyser (decomposition of a complex signal into its frequency components); a range of discrete 'channels' process visual information within a restricted range of spatial frequencies (Ginsburg and Hendee, 1997). These information channels represent the activity of the different cells and filter information about the visual target. Contrast sensitivity is more strongly predictive of visual detection tasks, including complex tasks where the viewer has to discriminate information in the optic array. The Contrast Threshold Function (CTF) is a measure of the minimum contrast needed for images to become distinguishable. Each human individual has a different CTF curve. In general, higher amounts of minimum contrast are needed at extremely low and extremely high spatial frequencies.

A contrast sensitivity (or response) function is created by plotting contrast sensitivity as a function of spatial frequency (Ginsburg and Hendee, 1997). The result is a curve (Figure 2.3) which describes a range of visible stimuli or 'window of visibility' (Girod, 1992).

The region below the curve in Figure 2.3 is the region of detectable contrast where objects are perceptible (where more contrast is needed for perception). Contrast Sensitivity decreases for spatial frequencies above and below the approximately two to six cycles per degree where Contrast Sensitivity is maximum.

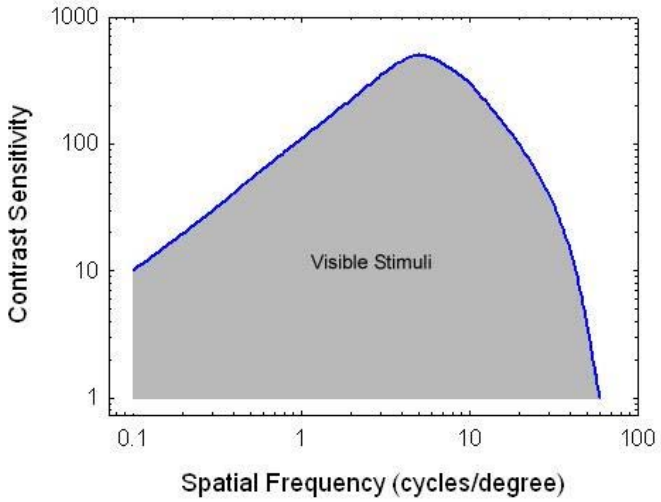


Figure 2.3 Contrast Sensitivity Curve

Girod (1992) described the spatio-temporal frequency response of the HVS in the design of a system where noise could be hidden at high spatial frequencies, recognising that the low spatial frequency response drop-off was less useful for this purpose. He described perceptually optimised coding schemes as ones which distribute the error in the image as evenly as possible but this approach does not consider the characteristics of the human retina and the fovea in particular. A wider consideration of psycho-visual phenomena in the design of systems, including a multi-channel model of human vision which includes the foveation response to 'important' image content, is the approach applied in this thesis (discussed further in Section 2.3.2).

Visual filtering mechanisms and contrast sensitivity, coupled with Fourier analysis theory, facilitate understanding of how visibility of objects changes as a function of size and contrast (Ginsburg and Hendee, 1997).

In addition to linear spatial and temporal filtering, retinal processing also includes multiple gain control mechanisms that maintain sensitivity to small changes in light intensity over space and time across a wide range of light conditions. Signals

transmitted from the retina to the brain must carry the maximum information about small spatial and temporal changes in light for large variations in light conditions.

2.3.2 The Spatial Response

The HVS is most sensitive to spatial frequency[5] of approximately 15 cycles per degree and falls off rapidly at both higher and lower frequencies, the maximum frequency being approximately 60 cycles per degree (a function of the size of the receptors rather than the properties of the lens and cornea). At high spatial frequencies the sum of outputs from rods degrades resolution. The low frequency response is a result of the process of lateral inhibition which effectively filters out low spatial frequencies. Although the observer is not aware of this limitation of the HVS in reproducing spatial information in a scene, it can lead to unusual effects in images, for example, the Mach Band Effect (Sharp and Philips, 1997).

Spatial resolution refers to the minimum image size of an ideal point source and therefore reflects an imaging system's ability to accurately reproduce the contrast available in an object (Hendee and Wells, 1997). The fovea (described in Section 2.1.1) is a specialised region of the eye responsible for high spatial resolution vision; essential for performing tasks like reading, driving and sports. Peripheral vision is low resolution (blurry) but adequate for processing large movement where detail is not required. Geisler and Perry (1998) found that savings in bandwidth can be obtained by matching the spatial resolution of transmitted images to the fall-off in spatial resolution of the HVS using image foveation techniques.

Image foveation is a method of modelling the human foveal response and producing image compression through reduction of spatial information. An example of the effect of the application of image foveation is given in Figure 2.4 which compares the original version of the standard test image 'Lena' (Figure 2.4(a)) with the same image which has been foveated from the central point of the image (Figure 2.4(b)). The size of the 'clear' foveal region is determined by the viewing distance and the degree of foveation

[5] Spatial frequency is described by the number of cycles per degree of visual angle. A fingernail at arm's length subtends about one degree of visual angle.

(blurring) is determined by the minimum contrast threshold set in the image foveation algorithm (described in Chapter 8). When viewed from a distance which places the clear region of the foveated image in the fovea, the blurring in the periphery is not detectable.

(a) Lena (original image)

(b) Lena (foveated image)

Figure 2.4 Lena Test Image (a) Original and (b) Foveated from the Central Point of the Image

Geisler and Perry (1998) identified the potential to exploit the decline of spatial resolution away from the point of gaze by the HVS. They identified two major limitations of the methods of image foveation for image compression which had been developed up to that period in time. The first of these is the appearance of blocking artefacts and motion aliasing[6] in peripheral regions, even at moderate degrees of foveation. The second limitation is the need for expensive and complicated real-time eye tracking to align the high resolution region of the display with the high resolution region of the eye (the fovea). Geisler and Perry (1998 and 1999) developed a multi-resolution pyramid

[6] Motion aliasing is the effect produced when continuous motion is represented by a series of samples.

model (using Laplacian pyramids with interpolation and blending to create a seamless transition between the resolution levels in the pyramid) for image foveation which reduced the effects of the first of these problems. It is important to note that detectable artefacts have a significant impact on the performance of these systems due to the high sensitivity of the HVS to spatial and temporal discontinuities. Although they recognised that a real-time system required knowledge of where the user was looking, to overcome this second limitation they identified potential applications (surveillance, telemedicine and teleconferencing) for user controlled foveation (using a simple pointing device) of localised regions of the image that required detailed inspection. Geisler and Perry (1999) obtained a reduction in bandwidth requirements by a factor of three for the 'News' test video sequence and a decrease in the data transmission rate or an increase in frame rate (by a factor of three) at a fixed data transmission rate. They reported transmission and coding benefits but their work was not presented in sufficient detail for direct replication of the application of the method or subjective quality results. The algorithm was designed for image compression and the demonstration (at http://www.svi.cps.utexas.edu) produced some visible artefacts and was limited to the resolution of the HVS (Geisler and Perry, 2002). Kleinfelder (1999) implemented the image foveation algorithm of Geisler and Perry (1998), using pixel value averaging rather than their multiresolution pyramid and interpolation methods, for application in an image sensor for 'vision enhancing' camera systems. Geisler and Perry (2002) developed their image foveation algorithm for application in the development of gaze contingent displays for controlling retinal stimulation for research on the roles of central and peripheral vision during the performance of complex tasks (for example, visual search and reading) and understanding of eye disease conditions. The determination of the point of foveation (gaze location, described in Chapter 6) and the development of methods of foveating moving video images (Chapter 8) were significant contributions in this thesis.

2.3.3 The Temporal Response

The eye responds to frequencies of up to 60 cycles per degree and a maximum rate of approximately 60 cycles per second and therefore a maximum of 5.5×10^{10} samples per second could be available for sending to the cortex, exceeding the bandwidth available in the optic nerve (Sharp and Philips, 1997). In other words, the temporal contrast

sensitivity of the HVS declines at high frequencies creating a temporal resolution cut-off at approximately 60Hz. The temporal response of the eye determines whether the eye sees a single flash of light. It depends on the number of quanta absorbed and the length of time the stimulus lasts. The maximum time over which the effect of quanta can be integrated is known as the critical duration of vision (t) and is approximately 0.1 second. The actual value depends on the luminance of light (B) according to Bloch's Law (Equation 2.1).

$$B \times t = \text{constant} \qquad \textit{Equation 2.1}$$

If the stimulus is flickering, rather than a single flash then, at a certain frequency, the eye sees steady light. This value is the Critical Flicker Frequency or Critical Fusion Frequency. The maximum flicker rate detectable by the eye is approximately 60 cycles per second (Sharp and Philips, 1997). This effect is used in cine (video) presentation of images.

Motion-compensated interpolation filtering by the eye, evident in smooth pursuit eye movements (Section 2.2.1), means that temporal frequency components can be perceived (Girod, 1992). This suggests that blurring of moving image content is perceived by the human eye. However, this is not the case for motion which is not tracked (Section 2.4) and so potential savings are possible for peripheral perception of motion in video images. The temporal response of the human eye is discussed in the context of motion perception in the following section of this thesis.

2.4 The Perception of Motion

Motion perception is the process of inferring the speed and direction of objects that move in a visual scene, given some visual input. While this process may appear simple to the observer, it presents a challenge from a computational perspective and is not fully defined in terms of neural processing. Motion perception theory is described in terms of first-order and second-order motion sensors in the HVS.

First-order motion sensors are stimulated by objects in which the moving contour is defined by luminance. A visual object in a scene may be distinguished as having a difference in luminance from its surroundings. When an object moves, its motion can

be detected by a relatively simple motion sensor which detects a change in luminance at one point on the retina and correlates it with a delayed change in luminance at a neighbouring point on the retina. These luminance-based or 'first-order' motion perception sensors have been specifically referred to as Reichardt detectors (Reichardt, 1969), motion-energy sensors (Adelson and Bergen, 1985) and Elaborated Reichardt Detectors (van Santen and Sperling, 1985). It is postulated that they detect motion by spatio-temporal correlation and provide a simple model for how the HVS detects motion. This model suffers from a phenomenon known as the 'aperture problem'. The aperture problem is such that each neuron in the visual system is sensitive to visual input in only a small part of the human visual field. Its scope is restricted to the visual field through a small window or aperture. This means that each neuron can only detect motion perpendicular to the orientation of the contour that is moving. The motion direction of a contour is ambiguous, because the motion component parallel to the line cannot be inferred from the visual input and therefore a variety of contours moving at different speeds will cause identical responses in a motion sensitive neuron in the HVS. Further processing is required to disambiguate motion direction. One way in which the aperture problem can be solved is by accessing information from knowledge/experience of the world. Since individual neurons respond to motion that occurs locally within their receptive field, each local motion detecting neuron will suffer from the aperture problem and so the estimates from many neurons need to be integrated into a global motion estimate.

Second-order motion perception sensors are stimulated by objects in which the moving contour is defined by contrast, texture, flicker or some other quality that does not result in an increase in motion energy in the Fourier spectrum (an invertible integral transform of one function into another) of the stimulus (Chubb and Sperling, 1988; Cavanagh and Mather, 1989). Nishida, Ledgeway and Edwards (1997) suggested that early processing of first-order and second-order motion is carried out by separate pathways. Second-order mechanisms have poorer temporal resolution and are low-pass (permit the passing of low frequencies and attenuate frequencies higher than a threshold frequency) in terms of the range of spatial frequencies that they respond to. Second-order motion produces a weaker motion after-effect. A motion after-effect is a visual illusion perceived after watching a moving visual object and then immediately looking at a stationary object. The effect is that the stationary object appears to move slightly in

the opposite direction to the original moving object. Motion after-effect is thought to be produced as a result of 'motion adaptation' (Ledgeway and Smith, 1994).

Flicker and movement sensitivity associated with temporal change do not follow the general rules of declining abilities of the fovea. Improved peripheral performance of these variables has been demonstrated by Baker and Braddick (1985).

Cavanagh (1992) demonstrated two attention-based motion processes; a 'low level' or automatic process that signals motion even when there is no attention to the stimulus and another that is mediated by attention to visible features and provides accurate velocity judgments independent of the features being tracked.

The organisation of movement in the changing image that reaches the eye provides the HVS with a valuable source of information for environmental analysis. The HVS uses relative movement to locate the boundaries of physical objects in the environment (Hildreth, 1984). In complex situations, for example distinguishing shadows from real object boundaries, the HVS has access to tacit[7] knowledge (Stoner, Albright and Ramachandran, 1990).

The HVS can resolve form and motion simultaneously, unlike a camera system, so that blurring of tracked moving objects is not perceived (Burr and Ross, 1986). Motion blur and motion sharpening effects in moving images are applied in systems to take advantage of the limitations of the HVS.

Motion blur is used in the television and cinema industry to capture the perceptual effect of high speed motion. The motion blur effect applied in the moving images is designed to mimic the behaviour of the HVS by taking advantage of the retinal smear phenomenon (Zagier, 2001). Visual information is accumulated over an interval of approximately 125 milliseconds and is blurred or smeared during this period (Burr and Ross, 1986). When the eye tracks motion, using smooth pursuit eye movements, the retinal velocity may be lower than the original image velocity. If the eye is moving, a stationary object will appear to be smeared. If the eye does not track motion, the

[7] Tacit knowledge is based on understanding of the situation or task from previous experience.

motion smear is due to the image plane velocity. This means that motion blur can be effective in smoothing motion in scenes of high image acceleration, low predictability of motion and complex scene motion (that is, when it is very difficult for the eye to track motion) so that the velocity on the retina matches object velocity (Zagier, 2001). If the image presented to the eye matches the expected retinal image, no blurring is visible. This implies that in scenarios where image features are not tracked by the observer, image detail may not need to be presented in high spatial and/or temporal resolution. This was important in the design of a video communication system optimised for the capabilities and limitations of the HVS of the viewer.

Blurred edges look sharper when they are moving than when they are stationary. This 'motion sharpening' effect also occurs when blurred images are presented for a short time (8-24 milliseconds) than for longer durations (100-500 milliseconds) without motion (Georgeson and Hammett, 2002). This evidence suggests that artificial sharpening of edges in blurred moving images does not necessarily enhance the perceptual quality of these images.

2.5 Summary and Model of Visual Perception

Human factors which influence the perception of visual information are important in the design of video communication systems optimised to meet the visual requirements of deaf people using BSL. The review of human vision in this chapter is summarised in a five-layer model of vision developed in this thesis to represent the oculomotor, visual, cognitive, context and human factors and features which influence the visual perception of BSL communication (Figure 2.5).

At the centre of the five-layer model is the oculomotor component of the HVS which is essential to the study of visual perception during an active visual task. Saccades direct the fovea to important content (features, objects and faces) which must be selected from the vast array of visual information available to the human eye. Visual selection via eye movements (and compensatory head and body movements) enables the HVS to give priority to visual regions which require high resolution vision during a fixation and to track motion during smooth pursuit movement.

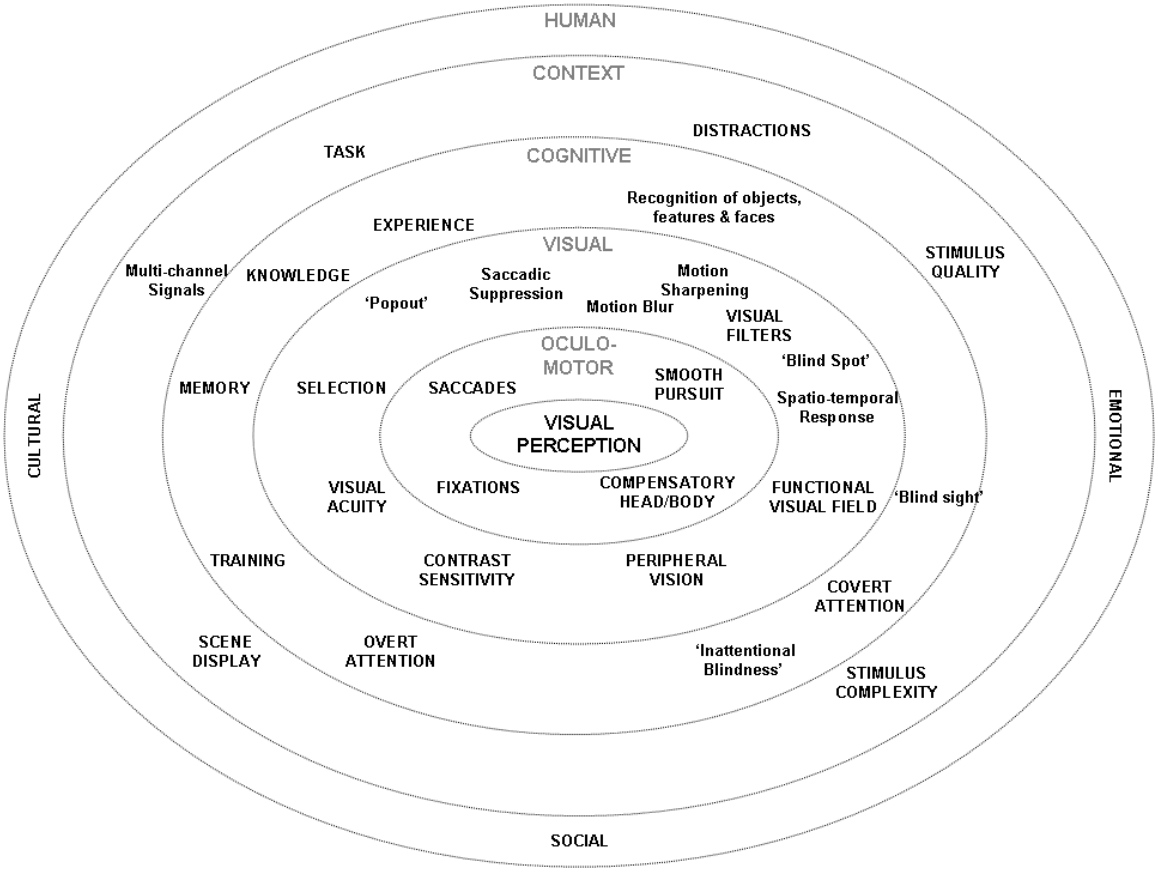


Figure 2.5 Five-Layer Model of Factors Influencing Visual Perception of BSL Communication

The next layer in the model of visual perception is the visual layer; this contains the HVS features and visual phenomena which enable efficient visual selection and signal processing. Since the capacity of the HVS is limited in terms of visual acuity, contrast sensitivity and the response to spatio-temporal change, it applies visual filters at each stage of visual processing to reduce the amount of visual information which must be processed. In spite of these HVS limitations, including the 'blind spot' (Section 2.1.1) where no vision is received, human vision appears smooth and uninterrupted. The fovea of the eye, which provides high visual acuity at the point of fixation, is directed to important spatial image content (especially faces) by rapid eye movements (saccades) which sample the visual environment. 'Saccadic suppression' (Section 2.2.1) is a phenomenon which inhibits visual processing of salient perceptual information during a saccade to reduce the amount of visual processing required. Visual acuity declines with

increasing eccentricity from the foveal centre. The temporal response does not conform to the same declining ability of the fovea and there is improved peripheral vision performance for motion detection. Visual phenomena, such as motion sharpening and motion blur effects (Section 2.4), aid the processing of motion in peripheral vision (Section 2.1.1), which is specialised for processing change in a large area of the scene (low frequency image components). In foveal vision, spatial selection is aided by 'popout' (Section 2.1.2). Individuals undertaking familiar complex tasks have enhanced ability to recognise important shapes (that is, visual signals such as hand and figure gestures in BSL) which 'popout' in the visual scene. The size of the functional visual field (Section 2.1.3) is also influenced by the processing capability of the HVS, foveal load and prior-experience of the task. Experienced sign language users have the ability to filter-out irrelevant information and maintain a wide functional visual field even when foveal load is high. Experience is a cognitive factor which has a very strong influence on the perception of visual information.

Efficient visual signal processing in the eye and brain precludes the necessity for an internal photographic representation of the outside world in the HVS and minimises the amount of information which must be processed in the nervous system. Efficient processing of visual information by the HVS depends on cognitive processes. The cognitive layer in the model of visual perception includes the analytical processes which make the connections between the visual context (that is, the nature of the task and the stimulus) with the visual and oculomotor processes in the HVS. BSL communication is a complex cognitive task requiring training, knowledge and experience which is accessed by the HVS to decode the visual signals of the language. The HVS accesses memory to enable efficient processing of recognised features, objects and faces (Section 2.1.3). Previous research has shown that deaf people have enhanced processing ability in the brain for face discrimination (Section 2.1.3.2), processing of motion (Section 2.4) and processing of peripheral information (Section 2.1.1). The cognitive layer also facilitates the important association between visual attention mechanisms and visual selection via eye movements (Section 2.2). Overt attention is by saccadic shift to a visual point of interest or importance and is the main form of attention in complex active tasks (such as BSL communication). Covert attention is directed away from the point of gaze to the periphery or to internal mental processes during passive vision. This implies that a study of eye movements with

respect to the user task (Chapter 6) is appropriate to determine the patterns of visual attention and visual regions-of-importance for the active task of BSL communication. The importance of analytical processing of visual information is also evident in the phenomena of 'inattentional blindness' (Section 2.1.3.2) and 'blind sight' (Section 2.1.2). 'Inattentional blindness' is the effect of not paying attention to different parts of the scene during a visual task so that some features or objects in the scene are not observed. 'Blind sight' is the ability of cortically blind subjects to respond to colour and motion (but not brightness) features of visual stimuli, providing evidence that the perception of different stimulus attributes is a function of different visual areas of the brain and that overall visual perception is analytical. What we perceive is not merely what we 'see'. This is fundamental to the study of visual perception of a complex task such as BSL communication and provides the rationale for a study of eye movements articulated in the context of the cognitive processes associated with it.

The context layer contains the critical factors related to the stimulus and the visual task which influence visual perception. The representation, quality and complexity of the stimulus influences the way in which it is processed by the HVS. The complexity of the task (including the multi-channel visual signals of BSL communication) and the presence of any cognitive or visual distractions also have an impact on visual perception of the scene. The visual task and the representation and quality of the visual stimulus are of critical importance in this thesis which aims to improve the perceived quality of video communication systems for BSL communication.

The outer layer of the model includes the cultural, emotional and social factors influencing visual perception. Cultural factors have been demonstrated to influence eye movements in natural scene perception (Section 2.2.4). Emotional and social factors influence visual perception in terms of the perception of visual quality (Section 3.3).

The following chapter (Chapter 3) considers the nature of the visual task (BSL communication) in more detail and the specific quality requirements and visual behaviour mechanisms influencing the perception of quality of video communication of BSL.

Chapter 3: British Sign Language Communication

This chapter explores the representation, diversity and complex visual components of BSL, the impact of these on visual behaviour and the need for optimised video communications. The technology currently available for video communication, the quality of service requirements and the perceptual mechanisms for BSL communication are discussed in the context of the development of an optimised system for deaf people.

3.1 British Sign Language

British Sign Language (BSL) is the most widely used method of signed communication in the United Kingdom, with an estimated 50,000 people using BSL as their first/preferred language (RNID, 2004). An understanding of how sign language is communicated is important in the design and development of a visual (video) communication system for deaf people using BSL.

The Royal National Institute for the Deaf identified three common misconceptions about sign languages which are relevant to this thesis (RNID, 2004). These misconceptions are that:

1. Sign language is simply a collection of gestures.
2. Sign language is a universal international language.
3. Sign languages are simply manual versions of the spoken language of that country.

These misconceptions are addressed in the following sections which explore the representation, diversity and complexity of the language.

3.1.1 The Visual Representation of Sign Language

Sign language (including BSL) is more than a collection of gestures. Sign languages are highly structured, linguistically complete, natural language systems that express vocabulary and grammar visually and spatially using a complex combination of facial expressions and gestures (such as eyebrow movements, eye blinks and mouth/lip shapes), hand and body movements, and finger-spelling that can change in space and

time (Stokoe, 2001). This means that the complexity of the language cannot be easily represented by a single written or spoken word or by a static image equivalent (Fels, Richards, Hardman, Soudain and Silverman, 2004). Video communication of sign language is therefore important in the equal provision of services for deaf people. This provides the rationale for the design of a video communication system for BSL communication.

3.1.2 The Diversity of Sign Languages

Sign language is not a universal international language. Sign languages are as diverse as spoken languages. Deaf people in different countries use different sign languages (for example, BSL in the UK, ISL (Irish Sign Language), ASL (American Sign Language), Auslan (Australian Sign Language) and Nihon Syuwa (NS) in Japan). Some countries have similar language structures, for example BSL and Auslan are similar. Auslan developed from the sign language of early immigrants from the UK. Similarly ISL and ASL have much in common. Sign languages continue to develop in the same way as spoken languages. Within each country there are also regional differences in the language, similar to dialects, where local variations of signs have developed. For example, in the UK, the author observed significant variations in the signs used in Aberdeen, Glasgow and Bristol. This fact was important in the design of BSL video material used in this research. It was important that all BSL video material used in the experimental work conducted in this thesis used the same regional version of BSL as the subjects viewing them so that there were no barriers to understanding the content apart from the perceived quality of the video images.

3.1.3 The Complexity of Sign Languages

Sign languages are not simply manual versions of the spoken language of that country. Sign language has its own grammar and syntax. BSL was officially recognised as a language in its own right by the UK Government in March 2003 (Scottish Executive, 2004). As well as being more than a manual representation of spoken language, sign language is more than a manual language. Non-manual signs are a very important component of sign language. The use of ‘facial behaviours’ in sign language (ASL) was examined by Corina, Bellugi and Reilly (1999) in fMRI (functional Magnetic Resonance Imaging) studies of the brain of deaf signers. Although the research described by these

authors was principally aimed at exploring the brain activity involved in processing different types of facial behaviour in sign language, some highly relevant observations were made in relation to the importance of information conveyed by the face and the processing of important facial information by deaf sign language users. Facial expression is used in sign language in two ways: to convey affect (as with spoken language) and also to mark specific grammatical structures (for example relative clauses) which is unique to sign language. Corina *et al* (1999) found that affective facial expressions of deaf signers were mainly mediated by the right brain hemisphere and linguistic facial expression involved left brain hemisphere mediation. In deaf signers, hemisphere specialisation in the brain is influenced by the purposes the signals serve. This suggests that visual information is processed according to its classification. This is discussed in Section 2.1.3 in the context of the perception of objects and features and contributed to the understanding of how BSL information is perceived by deaf people.

An assumption that the hands are the most important means of communicating information in sign language has led to sign language research that concentrated on studies of the phonology, morphology and syntax of manual gestures and the design of video communication systems which gives equal priority to the face and hands in video images (discussed in Section 4.2). However, sign languages have evolved to convey information in simultaneous channels. A variety of specific facial signals used in BSL is produced to co-occur with manual signals. These are evident in BSL as multi-channel signs (CACDP, 2003). Non-manual signals (for example, eyebrow rise or fall, head tilt, mouth shape, lip/tongue protrusion, eye gaze shift and eye blink) are important in conveying linguistic information. Facial adverbs are used in combination with manual signs to give detailed meaning to the signs.

Corina *et al* (1999) reported evidence that experienced sign language users have enhanced facial discrimination ability due to increased reliance on facial expression for linguistic contrasts. In an earlier study by Corina (1989), it was found that deaf signers exhibited specialised brain activity for the processing of facial signals according to their linguistic function. Deaf signers develop special abilities for perceiving distinctions relevant to sign language, particularly facial signals which are known to be very important for accurate sign language communication between deaf people (Emmory, McCullough and Brentari, 2003).

The importance of facial expression in sign language systems was highlighted in a study of visual behaviour by Patricia Siple (Siple, 1978). She stated that signers displayed a fixed pattern of viewing during sign language communication. She observed that, during a sign language conversation, the subject tended to look mainly at the signer's face with some small visual excursions around the face. This was attributed to the fact that the face provided important clues regarding the meaning of signs, even in a language that is gestural and where the hands have a significant role. This is explored in more detail in the context of the Human Visual System (HVS) in Section 3.4.

The complexity of the language, in terms of the importance of facial expression and the combinations and relative positions of gestures used to convey meaning, is illustrated in Figures 3.1 and 3.2. Figure 3.1 illustrates the use of mouth/lip shapes and gestures towards and away from the face of the signer. In this twelve-frame extract from a sequence ('Lisa Family') in which Lisa is introducing her family, she is explaining that her brother "has two children". In the first six frames (top row of Figure 3.1) the importance of the number of children is emphasised by the fact that she brings her right hand close to her face so that her mouth/lip shape and two-finger gesture (for the number two) can be observed together. In the next six frames (lower row of Figure 3.1), the sign for "children" starts close to the face and moves rapidly away from the face towards the end of the sign. The shape and speed of delivery of the sign do not require observation in fine detail and so it can be observed in peripheral vision while the detail of facial expression and lip/mouth shape, occurring at the same time, can be viewed during foveal observation of the face. Figure 3.2 illustrates the use of finger spelling to spell out the letters of Lisa's hearing dog's name, "Bran" in the 'Lisa Introduction' sequence. The finger-spelling of the name 'Bran' begins in frame number 128 and is completed in frame 153 (delivered in approximately one second in this sequence recorded at 25 frames per second). Single frames from this sequence are given in Figure 3.2 to distinguish the face and hand signals for each letter. Experienced BSL users are able to 'read' the rapid finger movements and the spatially detailed facial expression associated with finger spelling by directing their gaze to the face of the signer and attending to the hand shapes in peripheral vision.

The static frames in these illustrations do not truly represent the characteristics of the language evident in the moving image (as previously discussed in Section 3.1.1). In addition to providing motivation for improving video communication systems for deaf people using BSL, this also highlights the importance of studying the visual response to BSL video image motion (previously discussed in Chapter 2).

The following section reviews current video communication systems and identifies the opportunities and limitations of these for BSL communication.

Figure 3.1 Frames 252 to 263 from the 'Lisa Family' Sequence

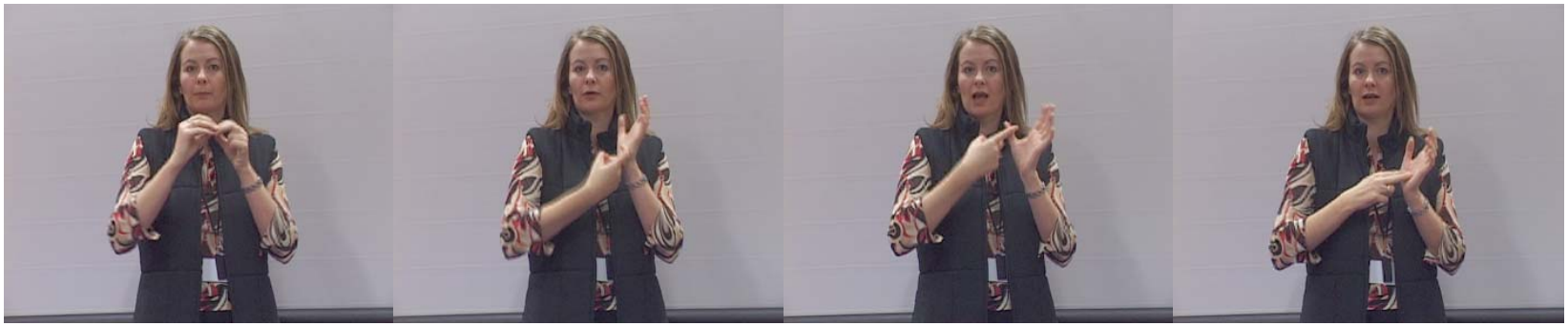


Figure 3.2 Frame 132 (‘B’), 138 (‘R’), 143 (‘A’) and 153 (‘N’) from the ‘Lisa Introduction’ Sequence

3.2 Video Telephony

Video telephony enables remote visual communication via terminals capable of transmitting video images (and audio signals). Video telephony is regarded as an enabling technology for users with hearing loss (providing an extra communication facility for lip reading and interpretation of facial expression) and in particular for those people who are pre-lingually deaf and use sign language as their preferred means of communication.

Telecommunications for video telephony ranges from standard analogue lines to multiple ISDN and high-speed digital links. The greater the available bandwidth the more (better quality) information can be sent and received. The quality of service enabled through the video telephony equipment and telecommunications facility is a trade-off between bandwidth and video (and audio) quality. Quality characteristics important to people who are deaf/partially deaf include high fidelity audio (to improve reception for those who are hard of hearing), synchronization of audio and video (to facilitate speech reading) and image quality for sign language users. Factors affecting image quality, which are of particular relevance to this thesis, are image size, image resolution (number of pixels) to improve the visibility of the image and the frame rate (number of frames per second) for motion perception.

Video telephony is covered by a range of ITU (International Telecommunications Union) standards which allow terminals and systems to communicate. These include standards for video (and audio) coding, interoperability between different types of equipment (for example high speed multimedia and low speed single medium) and text transmission (for written communication in the spoken language of the country of use) as part of a multimedia stream. It should be noted that text and audio content provision by video telephony are not considered in detail in this thesis as the main focus of the research is improving video quality for BSL users.

There are two main options for video telephony applications: a dedicated stand-alone terminal (video telephone) or a multimedia terminal (PC-based system) with video communication software. Many video telephone and video conferencing systems are aimed principally at business users; a very small number of vendors design systems

specifically for sign language users and so the choice of systems is very limited for the specific needs of this group of users.

This section describes the main options and services, identifies quality of service requirements and evaluates fitness-for–purpose of video telephony systems for sign language communication.

3.2.1 Video Telephones

Video telephones are generally stand-alone systems which include a video processor, video CODEC (encoder and decoder) and normally a camera and display screen. This option is relatively expensive to purchase but is easy to set up as it does not require hardware or software installations. Another advantage of this type of system is communication using a dedicated standard or high-speed circuit and so performance is not affected by network congestion. Major international suppliers of video telephones aimed at deaf users include Sorenson Communications (http://www.sorenson.com) and D-Link Systems (http://www.dlink.com). In addition to the purchase cost of these systems, important features for sign language users include screen/image display size, image resolution and video frame rate. Compatibility with other video telephones, Video Relay Services, the latest video CODEC standards (for improved picture quality at low bit rates) and broadband IP networks is also important for sign language communication. Displays range from small LCD screens on a stand-alone device to video telephone units requiring output to a TV or PC. Resolution is limited to QCIF (Section 4.1.4) on some systems and typical frame rates are up to 15 frames per second.

3.2.2 PC-Based Systems

A multimedia PC terminal provides a less expensive video telephony option for BSL users. This option requires additional installations of video communication software (for example, Microsoft NetMeeting) and hardware devices (for example, camera and microphone). Multimedia PC (and IP telephone) communication using the Internet means that the availability and quality of connection and image/audio quality varies considerably. Frame rates are typically around 15 frames per second and up to 30 frames per second is only achieved at high bandwidths (over 300 kbps).

3.2.3 Video Relay Services

Advances in Internet video communication have led to the development of important services for deaf people. Video Remote Interpreting (VRI) is the use of a remote Interpreter through a video connection when two people are together and they need an Interpreter. Video Relay Service (VRS) is a form of Telecommunications Relay Service (TRS) that enables sign language (including BSL in the UK) users to communicate with voice telephone users through video equipment, rather than through typed text (Figure 3.3). Video equipment links the VRS user with a Communications Assistant (CA) so that the VRS user and the CA can see and communicate with each other in signed conversation. The VRS caller, using a television or a computer with a video camera device and (generally) a broadband (high speed) Internet connection, contacts a CA, who is a qualified sign language Interpreter. They communicate with each other in sign language through a video link. The CA places a telephone call to the party the VRS user wishes to call and relays the conversation back and forth between the parties, in sign language with the VRS user, and by voice with the called party. A voice telephone user can also initiate a VRS call by calling a VRS Centre. Because communication is in the preferred language of the user (BSL), the conversation between the VRS user and the CA is easier and flows much more quickly than a text-based TRS call and the demand for VRS is very high in the UK.

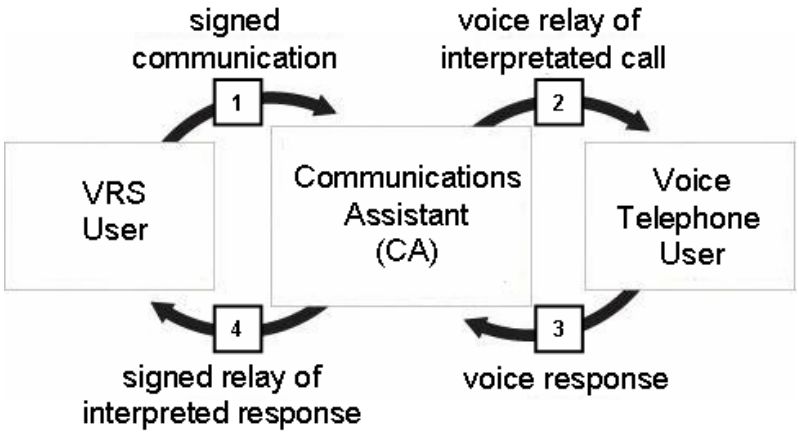


Figure 3.3 Video Relay Service (VRS) Communication

The first VRS centre in the United Kingdom was initiated in August 2005 in a joint venture between the BDA (British Deaf Association) and U.S. based CSD Inc. (Communication Service for the Deaf, 2005). The BDA-CSD VRS service was created to allow deaf consumers who use BSL to communicate with hearing people, using a D-Link videophone or PC with webcam and Internet connection, via a live video Interpreter who could speak to the hearing person using a standard telephone. The video Interpreter signs the telephone conversation with the sign language user. The Royal National Institute for the Deaf (RNID) VRS, RNID SignTalk, can be accessed via ISDN or broadband using video communication equipment such as a videophone or a computer with a webcam (RNID, 2006a). The user calls the RNID SignTalk VRS Call Centre and communicates in BSL; the Interpreter then makes a voice call and relays the conversation between the BSL user and the hearing person.

The BDA-CSD and RNID Sign Talk Video Relay Services were initiated to enable deaf and hard of hearing people in the United Kingdom to conduct fast seamless conversations during phone calls via qualified sign language Interpreters allowing the user's facial expressions and nuances of BSL to be portrayed as an important component of the language. Other similar projects have developed to attempt to meet the communication needs of deaf people at national and local level and for specific purposes. An example is the 'Significan't' Sign Video Project (http://www.significant-online.co.uk) which uses video communication technology to enable deaf people to communicate with Local Authority offices. In practice, the uptake of these services by the deaf community is limited by the cost of the equipment and service, availability of BSL Interpreters and the quality of the transmitted video communication. Deaf people have to make modifications to their natural expression of sign language to communicate within the constraints of small video image size, poor picture quality and transmission delays.

3.3 Quality of Service Requirements

Video telephony quality of service depends on the equipment and data capacity of the telecommunications systems. The video and audio signals are digitised and compressed, requiring significant computing power. The lines carrying the content must be capable of transmitting it sufficiently fast and without loss. Dedicated lines or circuit

connections can use the full bandwidth for this communication and so performance is generally better than that of the Internet which is a shared resource. Internet traffic causes delay and loss of data which can result in incomprehensible sign language. A number of researchers have investigated the quality requirements for video telephony in an attempt to define the minimum standards required for general communication tasks and for sign language communication.

In a comprehensive study of the fitness-for-purpose of video telephony, O'Malley, Brooks, Brundell, Hamnes, Heiestad, Heim, Hestnes, Heydari, Schliemann, Skjetne and Ulseth (1999) investigated the effects of video telephony quality of service (QoS) parameters (frame rate, video resolution, video delay and audio-visual synchrony) on task performance, communicative processes and user satisfaction. The tasks (T1 to T4) set in their experiments were designed to simulate 'real-world' scenarios relevant to broadband services. These were interviewing (T1), making social acquaintance (T2), instruction (T3) and joint problem solving (T4). Although not specifically looking at sign language communication, the tasks were susceptible to variations in quality of service parameters and produced variation in physiological factors involved in human communication relevant to sign language communication. The experimental design and methods of subjective and objective quality measurement, applied by O'Malley *et al* are also relevant to the study of sign language quality and are discussed in Chapter 5 of this thesis. Their results demonstrated the importance of the task and user expectation in subjective measures of video quality. Frame rate had a statistically significant effect on subjective measures of user satisfaction in acquaintance and instruction (but not interview and joint problem solving) tasks. Similarly, video resolution affected interview and joint problem solving (but not acquaintance) tasks. The researchers also found that prior experience of video telephony had an impact on user expectations of the service. New users were found to be less tolerant of low frame rate for instruction and more accepting of it for acquaintance tasks. Task performance (time and score for the task) and communication behaviour (number of interruptions, number and length of turns, frequency of individual and mutual gaze) were also measured as objective indicators of the fitness-for-purpose of video telephony. Frame rate had a statistically significant impact on task performance during interviewing and communication behaviour during instruction (but not on the other tasks). Low frame rates (below 15 frames per second) resulted in more verbal (audio) activity. Video resolution also affected task performance

in joint problem solving and communication behaviour in acquaintance, instruction and joint problem solving tasks. In these tasks lower resolution increased the length of turns and increased gaze frequencies. The results (summarised in Table 3.1) demonstrate the impact of a number of factors relevant to the perception of video quality for a BSL communication task. These include the effect of reduced frame rates and resolution on performance and communication behaviour and the level of prior experience of the viewer.

QoS Parameters	Subjective measures of satisfaction and efficiency				Objective measures of task performance				Objective measures of communication behaviour			
	T1	T2	T3	T4	T1	T2	T3	T4	T1	T2	T3	T4
Frame Rate		√	√		√						√	
Video Resolution	√		√	√				√		√	√	√
Audio-Video Synchrony									√	√		√
Video Delay							√				√	

Table 3.1 Effect of QoS Parameters on Task[8] Performance (O'Malley et al, 1999)

Woelders, Frowein, Nielsen, Questa and Sandini (1997) conducted a study to evaluate the picture quality requirements for speech-reading, finger-spelling and sign language in terms of frame rate and spatial resolution. A retina-like sensor was implemented in a video phone camera to produce high resolution in the central part of the image and degrading resolution in the peripheral region of the image. The purpose of this was to mimic the foveal response of the retina of the eye and reduce the number of pixels required to represent the picture with the aim of allowing a higher frame rate on standard telephone lines. The conclusion of this work was that speech reading was affected by frame rate (requiring at least 15 frames per second), finger-spelling and sign language were not affected by frame rate (tested at 10 and 15 frames per second) and none of the communication modes was affected by spatial resolution (tested at 6000 and 8000 pixels per frame). Edge distortion at the ring boundaries of the sensor

[8] Tasks: T1 - Interview, T2 - Acquaintance, T3 - Instruction and T4 - Joint problem solving.

was evident in the images presented and the range of parameters tested was limited. However, this work demonstrated that spatial quality requirements could be reduced without a reduction in perceived quality by mimicking the foveal response of the HVS. This was important in the design of a video communication system optimised for HVS capabilities and limitations (Section 4.2).

Hellström (1997) defined a quality profile for video communication of sign language. The profile was constructed based on measurements and observations using nine different video sequences transmitted through a pair of videophones. The tests were designed as a tool for judging the usefulness of a videophone for sign language communication. The measured quality characteristics provided useful guidelines for minimum acceptable standards of picture update rate (12-20 frames per second for smooth movement reproduction), static resolution (QCIF for acceptable communication; CIF required for good communication and essential for group discussions), motion details (fingers, eyes and mouth should be distinguishable, blurred fingers in motion are acceptable) and picture delay (less than 1.2 seconds for acceptability, good at below 0.5 seconds). Other factors affecting quality were lighting, background (plain background recommended) and a visual alerting system to ease initiation and turn-taking in communication.

Hellström and other researchers (ITU Study Group 16, 1998) identified the quality characteristics required from a system for real-time person-to-person conversation in sign language and lip-reading. The conclusion of this work was that the desirable performance goals were: 25-30 frames per second; CIF resolution; no more than a 0.4 second delay and the acceptance of 'occasional blur' during medium motion.

While these measurements provide a useful guide to minimum standards for individual quality components, they place a significant burden on video media applications. In addition, video content changes considerably during a natural sign language conversation and so quality requirements may change during the sequence. Variations in speed and size of manual and non-manual signs mean that an overall subjective view of quality in a natural situation may provide a more useful indicator of quality for the user. The use of a limited range of media and artificial laboratory experiments was necessary for measuring controlled variables in the fitness-for-purpose study for specific tasks by O'Malley *et al* (1999). However, the perception of quality is likely to be

affected by many additional complex human factors which are interrelated and can not be measured satisfactorily in an experimental set-up. This implies that a useful measure of quality can only be obtained for a very specific 'real' visual communication task (such as sign language communication). In particular, the viewing behaviour of the user may also affect the perception of quality. For example, the study by Hellström *et al* (ITU Study Group 16, 1998) measured the temporal requirements of sign language communication by considering the representation of the fingers during finger-spelling but this assumes that the subject is looking directly at the fingers during a finger-spelling event. The representation of small detailed movements (including eye gaze position to indicate direction and eye blinks for punctuation) and finger-spelling required further research in the context of how these motions are viewed by a sign language user. The differences observed for different modes of communication (speech-reading, finger-spelling and sign language) by Woelders *et al* (1997) and the apparent reduction in quality requirements when their retina-like sensor was applied in the videophone camera also raise questions about the viewing behaviour of subjects and demonstrate the potential benefit of accounting for this in the design of a video communication system. The measurement of subjective quality (Chapter 5) and the consideration of the viewing behaviour of deaf participants were central themes for this thesis.

3.4 Viewing Behaviour of Deaf People Using Sign Language

Analysis of the viewing behaviour of deaf people watching sign language provides important information for the design of optimised video communication systems. The visual response to a stimulus can be measured in terms of eye movements, fixations and scan paths (Chapter 2). How deaf people sample the visual environment and whether there is a characteristic pattern of viewing behaviour for recognised stimuli were important questions for further investigation in the primary research for this thesis. However, it has been found that the specific nature of the task (sign language communication in this case) and attention mechanisms (particularly the strength of the relationship between eye gaze and visual attention for this group of users) were also important factors to be considered in the study of viewing behaviour of deaf people.

Conscious visual perception during a complex task, such as understanding sign language information for personal communication, requires overt attention to the point

of regard (Section 2.5). However, the complexity of sign language communication including the use of manual, non-manual and multi-channel signs (Section 3.1) also requires covert attention to BSL action in the peripheral field of vision. It may be the case that visual information gathered from the peripheral view of the scene, mediated by the cognitive and contextual factors described in the model of visual perception developed for this thesis (Figure 2.5), is sufficient for decoding the sign language message conveyed to the viewer. Understanding where the viewer is looking is very important. Analysis of this data in the context of the nature and visual processes of the visual task is crucial when drawing conclusions about viewing behaviour and the quality requirements for display of that information.

Visual behaviour during the task of sending and receiving sign language signals in a sign language conversation was explored by Patricia Siple (1978) in her observations of sign language conversations[9]. Siple proposed that, since sign language is received and initially processed by the visual system, then the rules for forming signs would be constrained by the limits of that system. She observed that subjects viewing sign language looked at the face, with small excursions around the face, of the signer. This behaviour demonstrated the importance of the face in giving clues to the meaning of gestures. Her paper studied the development of the sign system to maximise the information that the eye can gather. In sign language production, small detailed motions were observed to occur in and around the face and upper body region where the receiver (looking at the signer's face) could observe gestures in high acuity. Large, less detailed gestures are produced in the peripheral region of view and are therefore captured by the receiver at low visual acuity. These large motions tend to be in the vertical and horizontal axes where acuity is greater than for other orientations. Siple also described the exploitation of redundancy to further maximise the information that could be conveyed in the peripheral region of view. The conclusion of the study by Siple is that efficient communication of sign language between deaf people had developed within the constraints of the Human Visual System. The same limitations may therefore be exploited in the design of video communication systems for deaf

[9] Siple (1978) made physical observations of signers. Her conclusions were not based on eye movement tracking.

people using BSL. Eye movement tracking studies of deaf people watching sign language video were conducted in this research (Chapter 6) to provide evidence for the observations made by Siple (1978) and inform the design of a perceptually optimised video coding system which gives priority to the visually 'important' regions of the video image.

Gaze patterns and facial expressions provide an extremely important and rich set of social signals that help to regulate conversation as well as expressing intimacy and social control (Bruce, Green and Georgeson, 2003). The timing and direction of such facial gestures are crucial for their interpretation. Gaze pattern and duration of gaze give clues about the attention and communication cues (taking turns) during a conversation (spoken or signed). Humans are accurate at detecting gaze direction (up to one minute of arc of visual angle) so that they can not easily ignore cues to direction from eye and head movements which influence the speed of response to verbal directions (Bruce, Green and Georgeson, 2003).

In real-time video communication applications (such as video conferencing) mutual gaze is another important aspect of visual communication (Grayson and Monk, 2003). Conventional systems do not support natural mutual gaze due to vertical disparity between the camera and the image of the receiver's eyes. Grayson and Monk (2003) argued that mutual gaze information may be interpreted by the brain. Eye contact is known to be important for sign language communication and so gaze awareness (although beyond the scope of this research) may be a factor for investigation in the future development of real-time video communication systems for deaf users.

3.5 Summary

BSL is highly structured and diverse with international, regional and local variations. It expresses vocabulary and grammar visually and spatially using a wide range of complex and interrelated manual and non-manual gestures. As such, the language can not easily be represented in written or static image form and so video communication systems are regarded as an enabling technology for deaf people. An increase in the range of video services and available bandwidth capacity has led to increasing demand for video communication for deaf people for interpersonal communication and for access to work and vital services (including video relay and interpreting). Service entry

barriers include the high cost of equipment and ‘premium’ tariffs for high bandwidth services. Barriers to effective communication include poor image quality, transmission delay and small picture size which require the user to modify their natural expression of sign language. Optimal design of the systems which support these services requires a trade-off between available network bandwidth and video image quality.

Video coding techniques have developed to improve perceived quality and reduce bandwidth requirements by taking advantage of the limits of visual acuity and contrast sensitivity. It is postulated that further quality gains can be made at low bit rates by optimising video communication systems to meet the specific task requirements, quality of service requirements and the visual behaviour mechanisms of BSL users. Current video compression standards and options for optimisation are evaluated in the following chapter.

Chapter 4: Video Coding

The nature of BSL communication, quality of service requirements and the visual response of deaf people to BSL communication are important factors in the design of video communication systems optimised to provide best quality for the user. This chapter evaluates current video compression methods and available methods for optimising video coding.

4.1 Video Compression

A video sequence is a set of continuous still images (frames) captured at a particular frame rate. Each frame consists of a number of pixels (picture units), depending on the video frame format. The following sections describe digital video formats, the rationale for compression and the process and standards for video coding.

4.1.1 Video Frame Formats

The video frame format defines the pixel resolution of the image described as the number of pixels per line by the number of lines per picture. Each video frame is composed of one luminance component (Y), which specifies the intensity level of the pixels, and two chrominance components (Cb and Cr [10]) which indicate the corresponding colour difference information in the frame. The resolution of each chrominance component is equal to half the resolution for the luminance component. The rationale for this is that the human eye is less sensitive to colour detail since the retina of the eye has fewer colour sensors (cones) than luminance sensors (rods) and so the sensitivity to luminance stimuli is greater than that for chromatic stimuli (as described in Section 2.1). This is important in the context of the perception of image quality (Westland, Owens, Cheung and Paterson-Stephens, 2006). The standard formats are based on Common Intermediate Format (CIF). The pixel resolutions of the standard video frame formats for block-based coding (Section 4.1.3), used in video telephony applications, are shown in Table 4.1.

[10] Cb and Cr are the blue and red choma components.

Video Frame Format	Luminance Resolution	Chrominance Resolution
sub-QCIF (Sub-Quarter CIF)	128×96	64×48
QCIF (Quarter CIF)	176×144	88×72
CIF	352×288	176×144
4CIF	704×576	352×288
16CIF	1408×1152	704×576

Table 4.1 Frame Resolution of Intermediate Formats (Horizontal ×Vertical Pixels), Adapted from Sadka (2002) and Richardson (2002)

4.1.2 Rationale for Video Compression

Assuming that a video frame is digitised to represent each pixel, the number of bits required to represent each frame can be calculated. For example, for CIF format and eight-bit precision for each luminance and chrominance component, each picture would require 152 Kbytes (that is, ((352×288) + 2(176×144)) bytes). If the video frames were transmitted without compression, at 25 frames per second, the raw data rate for the video sequence would be approximately 30 Mbps and a one-minute clip would require approximately 225 megabytes of bandwidth. Consequently, digital video data need to be compressed prior to transmission to optimise the bandwidth requirements of video communication systems.

It could be argued that, with the falling cost and rising capacity of storage and transmission and the wide availability of broadband services in the UK, video compression efficiency is not a priority for modern communication systems. However, at low and high transmission bandwidths, an efficient video compression system provides significant performance advantages for visual communication (Richardson, 2003). Efficient compression offers improved perceptual quality within the constraints of increasing user demand and ‘premium’ commercial tariffs set for higher bandwidth limits. Enhanced coding efficiency can enable the coding of more video channels or higher quality video representations within existing digital transmission capacities (Wiegand, Sullivan, Bjontegaard and Luthra, 2003). Video telephony applications (including the full range of vital services such as the development of Video Relay

Services) for sign language users will continue to benefit from robust, efficient video coding optimised for the specific needs of the user. The development of the latest H.264/AVC Standard (Section 4.1.4) for video compression, by the ITU-T, VCEG (Video Coding Expert Group) and ISO/IEC MPEG organisations, represents the continuing advance in standard video coding technology in terms of coding efficiency and flexibility for effective use on the wide range of network and application types (Wiegand, Sullivan, Bjontegaard and Luthra, 2003).

4.1.3 Video CODEC

Video compression (coding) is the process of reducing digital video data to a smaller number of bits for storage and transmission. Video compression systems include a CODEC (enCOder/DECoder). The encoder transforms the video data into a compressed form and the decoder converts the compressed data into a representation of the original video. Video compression is achieved by removing redundant data. Video compression methods exploit temporal redundancy (between adjacent frames) and spatial redundancy (between neighbouring pixels) in video images.

Lossless compression is the process of removing statistical redundancy in data. In this type of system the decoder output (reconstructed data) is an exact copy of the original data and there is only a small degree of compression.

Lossy compression is applied to give the required compression ratios for video communication systems. In a lossy compression system subjective redundancy is removed (that is, data which can be removed without significant impact on the viewer's perception of visual quality) and the output reconstructed image is not identical to the source data. The performance of the HVS is limited due to its anatomical structure and neural processing capability (Section 2.1). Removal of subjective redundancy is based on knowledge of the sensitivities of the HVS to video image display in terms of colour, brightness, contrast and spatio-temporal resolution (Table 4.2). The implications of the limitations of the HVS, in terms of the relative sensitivity to these display characteristics, are that reductions in video image resolution can be made resulting in lower transmission bandwidth requirements without loss of perceived video quality by the user.

HVS Sensitivity to …	Design Implications …
luminance rather than chrominance detail	reduction of chrominance resolution
high contrast (large differences in luminance)	preservation of high contrast regions of the image (particularly at edges) rather than low contrast detail
low spatial frequencies (changes in luminance and chrominance over a large area)	removal of some of the higher frequencies (rapid changes in luminance in a small area of the image) while preserving edge detail.
persistent image features	reduction of temporally persistent artefacts
temporal change at less than 20Hz	optimisation of frame rates (20-30Hz) for the perception of 'smooth' motion

Table 4.2 Sensitivity of the HVS to Video Image Display and the Implications for Digital Video System Design (Adapted from Richardson, 2002)

Lossy compression achieves high compression rates at the expense of overall video quality (Richardson, 2003). A video communication system makes a trade-off between quality and the data rate achieved by compression. The measurement of video image quality is important for this trade-off decision (Chapter 5) in systems designed for different visual media applications.

A basic 'lossy' CODEC system processes each input video frame through a number of stages shown in Figure 4.1.

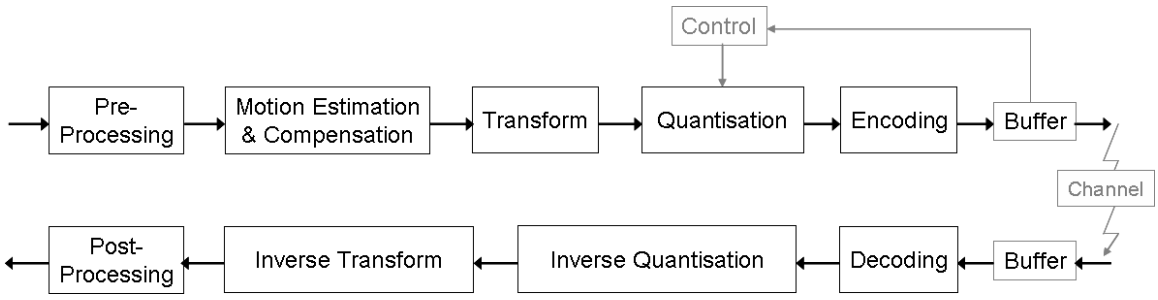


Figure 4.1 Block Diagram of the Processing Stages of a Video CODEC (Adapted from Sadka, 2002)

The processing stages of a video CODEC are described in the following sections. The design of the pre-processing, transform and quantisation stages are particularly important in the optimisation of a video CODEC since they have a direct impact on the nature and degree of compression.

4.1.3.1 Pre-Processing and Post-Processing

Pre-processing techniques (for example, noise filtering) may be applied to suppress or enhance particular features of input frames before encoding to improve the efficiency of the coding process (Sadka, 2002). Similarly, filters may be applied at the post-processing stage (for example, de-blocking, 'de-ringing' and/or edge enhancement filters) to improve the quality of the decoded images (Richardson, 2002).

In this thesis, pre-processing filters were applied to remove spatio-temporal data in image scene regions (selected based on studies of viewing behaviour) as a means of compressing the image data before encoding (Chapter 8).

4.1.3.2 Motion Estimation and Compensation

Improved compression performance can be achieved in a video CODEC by exploiting temporal redundancy between video frames by referencing neighbouring frames to predict and compensate for motion (inter-frame coding). This involves prediction of the current frame based on one or more previously transmitted frames and subtracting the prediction from the current frame to produce a 'residual frame' (Richardson, 2002). Performance is improved further in standard video CODECs by exploiting the differences in regions (blocks) of the video image, caused by changes due to motion, using motion estimation and compensation (Figure 4.2). A block (A), with dimensions m×n, in the current frame (normally a block of luminance samples) is compared with neighbouring regions of the previous reconstructed frame. Motion estimation is achieved by finding the best match to determine the neighbouring block (B) in the reference frame that gives the smallest residual block. A full search or one of a number of different fast search methods (for reduced complexity) may be applied to determine the 'best' match. The matching region in the reference frame, identified in this process, is subtracted from the current block (A - B) to achieve motion compensation.

The decoder carries out the same motion compensation operation to reconstruct the current video frame. This requires the encoder to transmit the location of the best matching blocks in the form of a set of Motion Vectors (MV in Figure 4.2). The decoder does not perform motion estimation as this information is transmitted in the coded bit stream.

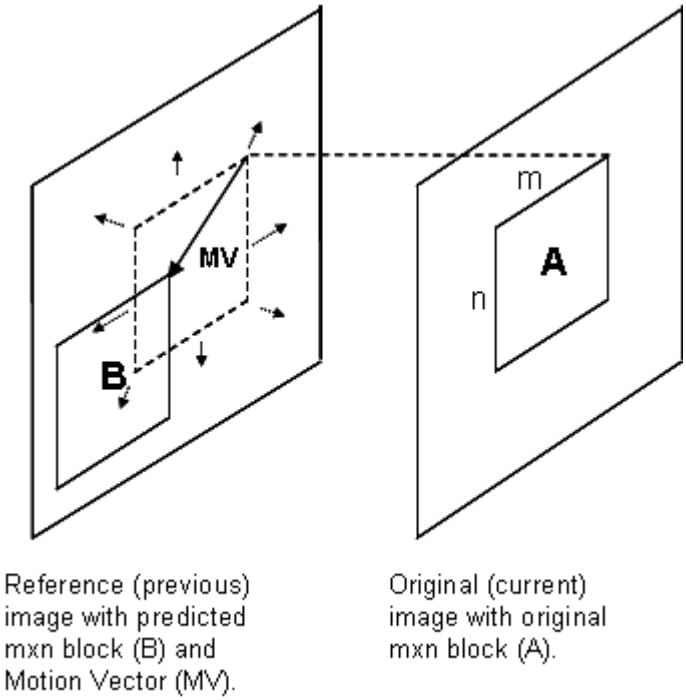


Figure 4.2 Block-Based Motion Estimation and Compensation

Motion estimation and compensation is relatively computationally intensive and the design of this stage has a significant impact on the overall compression performance and computational complexity of a video CODEC (Richardson, 2002). Although this stage contributes to the removal of redundant data, it does not have a significant impact on compression ratios.

4.1.3.3 Transform and Inverse Transform

Following the pre-processing and motion estimation stages, statistical redundancy in video frames is eliminated by mathematical transformation; for example, Discrete Cosine Transform (DCT) or Discrete Wavelet Transform (DWT) in the transform stage. No compression is achieved at this stage. The image data are separated into components of varying 'importance' to the appearance of the image but data are not removed at this point in the process (Sadka, 2002).

The DWT is usually applied to image sections or 'tiles' or to complete images and this method is applied in still image coding standards. Although not widely implemented for

motion video compression, wavelet transform is under research and development for video communication systems. The Dirac project (http://dirac.sourceforge.net) was initiated by the BBC (British Broadcasting Corporation) Research and Development department. The project aims to develop an open DWT-based CODEC in collaboration with the open source community as a potential alternative to proprietary or standard video compression systems.

The DCT is widely used in the current generation of video coding standards due to its effectiveness in transforming image data into a form which is easily compressed and its efficiency in terms of its implementation in hardware and software. The main advantage of the DCT is that it is optimal in terms of its compression capabilities based on human visual response mechanisms (Raff, 1997). The DCT is applied to blocks of data (discussed with reference to the different Standards in Section 4.1.4). The FDCT (Forward DCT) transforms a block of image samples (the spatial domain) into a block of transform coefficients (the transform domain) using a mathematical algorithm which was first proposed by Ahmed, Natarajan and Rao (1974). The transform represents each block of image samples as a weighted sum of 2-D cosine functions, referred to as DCT basis functions which are represented as DCT basis patterns in Figure 4.3.

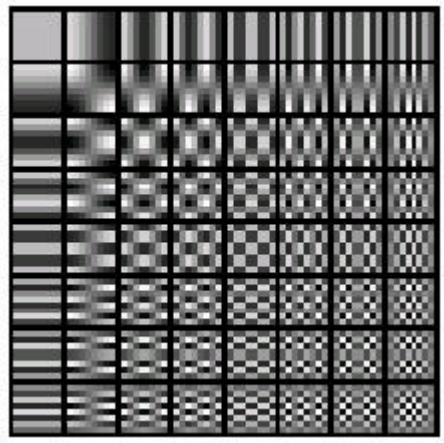


Figure 4.3 DCT Basis Patterns (Source: The MathWorks, http://www.mathworks.com)

The top-left basis pattern in Figure 4.3 is a uniform block and has the lowest spatial frequency. Horizontal spatial frequency increases moving to the right and vertical spatial frequency increases moving down towards the bottom-right block which has the highest horizontal and vertical spatial frequency. The selection of perceptually

'important' spatial image frequencies in the DCT is based on a general model of the HVS; that is, the eye is more sensitive to the information in the DCT coefficients that represent low frequencies (corresponding to large image features) than to those that represent high frequencies (corresponding to small image features). The DCT contributes to efficient compression by giving priority to the low spatial frequency coefficients and concentrating video image energy into a small number of coefficients as a result. This is illustrated in the example of a DCT transform of a block of pixels given in Figure 4.4(a). The transform of the video image data in the original block produces coefficients in the transformed block which are not evenly distributed. A few large coefficients are positioned in the upper left-hand corner of the block and so the DCT has reduced the spatial redundancies in the block and suppressed the correlations of the original pixels. The energy of the block is concentrated in the top left-hand section where lower frequency coefficients in the original block are located. The DC coefficient (top left) is the most significant coefficient in any block, corresponding to the average of the block pixels. Block transform video coding algorithms exploit HVS sensitivity to lower frequency DCT components by giving priority to the perceptually important DC coefficient of the block so that it is coded more accurately than the remaining high frequency (AC) coefficients.

Discrete Cosine Transform (Rao and Yip, 1990) of an 'N×N' block of image samples (or residual values after prediction) is given in Equation 4.1.

$$F_{x,y} = C(x)C(y)\sum_{i=0}^{N-1}\sum_{j=0}^{N-1} f_{i,j} \cdot \cos\left(\frac{(2i+1)x\pi}{2N}\right)\cos\left(\frac{(2j+1)y\pi}{2N}\right) \quad \textit{Equation 4.1}$$

Where,

$$C(n) = \begin{cases} \sqrt{\frac{1}{N}}, for \quad n = 0 \\ \sqrt{\frac{2}{N}}, for \quad n > 0 \end{cases}$$

$f_{i,j}$ represents the samples of input block where $f_{0,0}$ is the top left hand corner sample of the block. $F_{x,y}$ represents the coefficients of the transform block where $F_{0,0}$ (the 'DC' coefficient or the average) is at the top left corner of the block.

The transform is reversed in the IDCT (Inverse DCT). The IDCT transforms a set of coefficients into a set of image samples (Figure 4.4(c)). The significant coefficients are clustered around the DC components and so a good approximation of the original block is reconstructed from this small number of coefficients (Richardson, 2002).

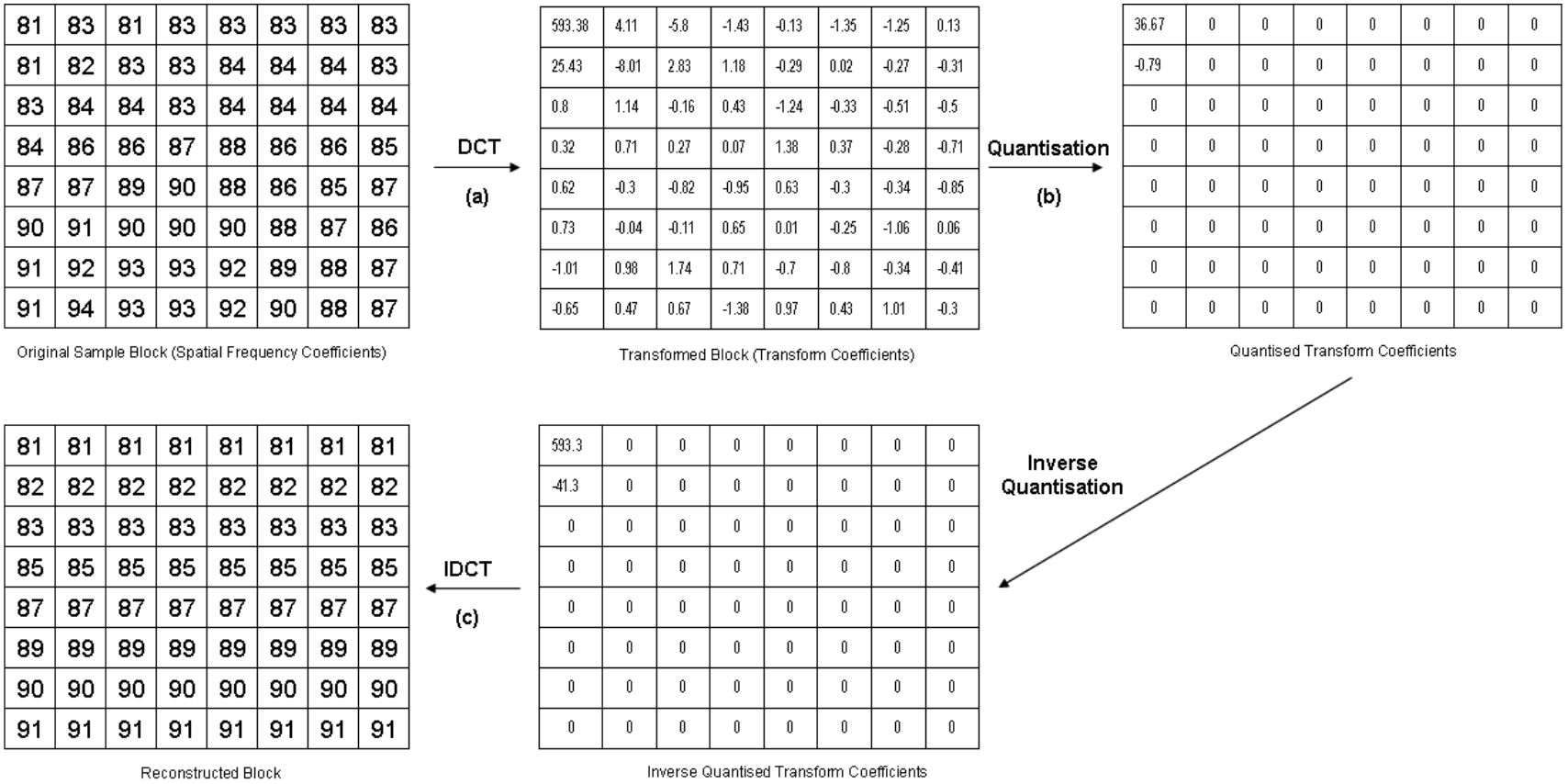


Figure 4.4 (a) DCT Transform of a Block of 8x8 Pixels, (b) Coarse Quantisation of Transform Block Coefficients, (c) Reconstructed Block Obtained from Inverse Quantisation and IDCT of the Transform Block Coefficients (Adapted from Sadka, 2002)

4.1.3.4 Quantisation and Inverse Quantisation

The next stage of the video compression process is quantisation. The quantiser controls the coding efficiency and quality of the reconstructed video sequence (Sadka, 2002). The quantisation stage removes information which does not have a significant impact on the appearance of the reconstructed image. This results in an irreversible loss of quality (lossy compression). The inverse quantisation process at the decoder converts the quantised values to the corresponding reconstructed values.

The quantiser maps the transformed coefficients (COF) to a smaller range of absolute values (Levels) of the quantised coefficients, depending on the level of quantisation, to reduce the number of bits required to code the block. The degree of coarseness of quantisation is controlled by the quantisation step size (which corresponds to two times the Quantisation Parameter (QP) set in the encoder). An example of a forward quantisation in the H.263 standard is given in Equation 4.2 (Sadka, 2002).

$$\text{Level} = (|\text{COF}| - \text{QP}/2) / (2 \times \text{QP}) \qquad \textit{Equation 4.2}$$

Fine quantisation (low QP value) leaves most of the coefficients in the quantised block and course quantisation (high QP value) removes all but the most significant coefficients (Figure 4.4(b)) thereby giving greater compression at the expense of greater loss in video image quality in this case. The QP value can be set to obtain the optimum balance between compression efficiency and video image quality for the application. The application of the Inverse Quantiser to the rescaled block and IDCT to the block of transformed coefficients produces a block which has lost information in the quantisation process but is similar to the original image block (Figure 4.4(c)).

4.1.3.5 Encoding and Decoding

After quantisation, the remaining significant transform coefficients are entropy encoded together with side information (such as headers and motion vectors) to form a compressed representation of the original video sequence. A typical image block will contain a minority of significant non-zero coefficients and a majority of zero coefficients after the DCT and quantisation stages (Richardson, 2002). The non-zero data are grouped together in sequence in a ‘zig-zag’ scan order from the lowest to highest

frequencies and can be represented as a series of (run, level) pairs indicating the number of zeros followed by the value (level) of the non-zero number. For example, (6, 12) represents 6 zeros followed by 12. Entropy coding, applying statistical compression algorithms (for example, Huffman and arithmetic coding), is applied to the run-level data. The frequently occurring pairs are represented with a short code and the less frequently occurring pairs with a longer code so that run-level data can be compressed into a small number of bits. The output of the entropy encoder is a sequence of binary codes representing the original image in a compressed form. The decoding process involves extracting the run-level symbols from the bit stream, converting them to a sequence of coefficients, reordering into a block of quantised coefficients, rescaling by inverse quantisation (lossy) and inverse transform to produce a reconstructed image where similarity with the original image depends on the quantisation level.

4.1.3.6 Rate Control Buffering

The bit rate generated by video coders is highly variable due to the temporal activity of video signals and the application of variable-length encoding. To regulate the output bit rate in real-time transmissions a buffer is applied between the encoder and recipient network (Sadka, 2002). A feedback control mechanism is used to regulate the encoding process according to buffer occupancy (Figure 4.1). Buffer-based rate control algorithms such as Scaleable Rate Control, and other rate control techniques, are important for the performance enhancement of CODECs but not considered in depth within the scope of this thesis.

4.1.4 Video Coding Standards

This section provides an overview of the development of international standards in video coding and highlights the importance of compression efficiency and optimisation of video communication systems to meet increasing user demand for services and quality.

Advances in digital transmission links and video coding technology have resulted in the development of a wide range of applications in the last twenty years. During this period world-wide commercial interest in video communications, such as video-on-demand, multimedia video database services, digital television and High Definition Television

(HDTV), led to the development of new international coding standards (Figure 4.5). The aim of international standardisation of video coding is to foster collaboration and the development of efficient digital representation of video signals (Sikora, 1999). The standardisation process typically involves identifying requirements for a specific application (or field of applications); competitive exploration and comparison of different algorithms; selection of a single basic technique; collaboration to refine the technique and produce a draft standard; validation of the standard by compliance testing and field trials and, finally, issue of the International Standard following any refinements arising from the validation process.

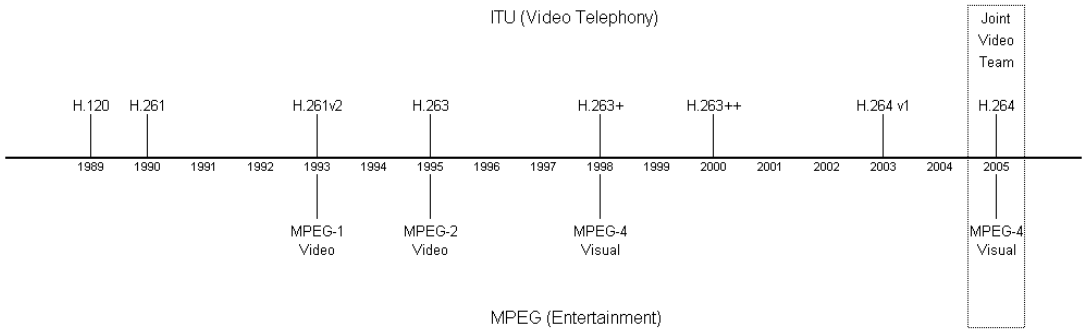


Figure 4.5 Video Coding Standards Publication Timeline

Study Group (SG) 15 of the CCITT (Comité Consultatif International Téléphonique et Télégraphique, known as the ITU - International Telecommunication Union since 1992) was the first international committee to start standardisation of video coding in 1980 and issued Recommendation H.120 for videoconferencing applications (CCITT, 1989). Recommendation H.120 did not achieve the goal of world-wide standardisation and this was one of the reasons why it was not a commercial success.

Advances in video compression research and technology led to the development of a new standard for videophone and videoconferencing applications at low bit rates. The H.261 video coding standard (ITU-T, 1993) was originally designed for transmission of CIF and QCIF frames over ISDN lines with data rate multiples of 64 kbps (CCITT, 1990). The data rate of the coding algorithm was designed to be able to operate at bit rates between 40 kbps and 2 Mbps. The basic processing unit of the H.261 design is the macroblock consisting of a 16×16 array of luminance samples and two

corresponding 8×8 arrays of chrominance samples. Inter-frame prediction removes temporal redundancy, with motion vectors to compensate for motion. Transform coding, using an 8×8 Discrete Cosine Transform (DCT), removes the spatial redundancy. Scalar quantisation is applied to round the transform coefficients to the appropriate precision, and the quantised transform coefficients are 'zig-zag' scanned and entropy coded (using a Run-Level variable-length code) to remove statistical redundancy. The H.261 standard specifies how video is decoded. Encoder designers could implement their own encoding algorithms, as long as the output was constrained to allow it to be decoded by any decoder conforming to the standard. Processing prior to encoding (pre-processing) or display (post-processing) was also left open for CODEC designers. A post-processing technique that became a key element of the most advanced H.261-based systems was the application of a de-blocking filter to reduce the appearance of block-shaped artefacts caused by block-based motion compensation and spatial transform. De-blocking filters were subsequently adopted as a feature of later standards. The refinements introduced in later standards have resulted in significant improvements in compression capability relative to the original H.261 design. H.261 has been made practically obsolete as a result although it is still used as a backward-compatibility mode in some video conferencing systems and for some types of Internet video.

The H.263 CODEC (ITU-T, 1995) was developed as a low bit rate coding solution for videoconferencing systems. It was initially designed to be implemented in H.324 (circuit-switched network videoconferencing and video telephony) based systems. It was later applied in H.323 (IP-based videoconferencing), H.320 (ISDN-based videoconferencing), RTSP (Real Time Streaming Protocol, for streaming media) and SIP (Session Initiation Protocol, for Internet conferencing) systems. H.263 was developed as an improvement on H.261 (for all bit rates) and the MPEG (Motion Picture Expert Group) standards developed by the ISO/IEC (International Standards Organisation/International Electrotechnical Commission). The original MPEG-1 (ISO/IEC, 1993) standard was designed for video compression for storage and playback on CD-ROM. The MPEG-2 (ISO/IEC, 1995) standard was developed for efficient coding of television-quality video with support for interlaced video, scalable coding (providing optional enhancement layers to improve decoded quality) and profiles and levels for controlling the functional capabilities of the application

(Richardson, 2002). H.263 was further enhanced in later versions of the standard; H.263+ (version 2, 1998) and H.263++ (version 3, 2000).

The development of the MPEG-4 (ISO/IEC, 1998) standard included many of the features of MPEG-1 and MPEG-2 with additional support for low bit-rate applications and object-based coding and a flexible set of coding tools for optional inclusion of additional features by developers (Richardson, 2003). However, it was unlikely that a developer would implement the entire MPEG-4 set of standards and so the standard included enhanced profiles and levels based on those defined in the previous standard (MPEG-2). These allowed a specific set of capabilities to be defined for a particular subset of applications. One of the most important enhancements in the MPEG-4 standard was the inclusion of object-based coding which allowed a video scene to be handled as a set of foreground and background objects (Video Objects) in the coding scheme, rather than merely as a series of frames (Richardson, 2002). This major development in coding standardisation led to possibilities for optimising video coding based on natural scene content (discussed further in Section 4.2).

The development of the H.264/AVC (Advanced Video Coding) standard was a joint venture between the ITU-T Video Coding Experts Group (VCEG) together with the ISO/IEC Moving Picture Experts Group (MPEG) collectively known as the Joint Video Team (JVT). The ITU-T H.264 standard (ITU-T, 2003) and the ISO/IEC MPEG-4 Part 10 standard (formally, ISO/IEC 14496-10) are technically identical. The H.264/AVC standard was designed to provide good video quality at bit rates substantially lower (that is, half or less) than would be required by MPEG-2, H.263 or MPEG-4 Visual without significant increase in complexity which would make the implementation impractical. The standard was designed to be applied to a wide variety of applications (for both low and high bit rates, and low and high resolution video) and to work on a wide variety of networks and systems. The development of the standard provided additional features at each stage of the coding process. These included, Variable Block-Size Motion Compensation (VBSMC) and new transform design features. VBSMC, ranging from 16×16 to 4×4, enables precise segmentation of moving regions. At the transform stage, features include adaptive encoder selection between 4×4 and 8×8 transform block sizes for exact-match integer transform. The special case of 4×4 integer transform allows precise placement of residual signals to reduce the

'ringing' effect encountered in earlier CODEC designs and the 8×8 spatial block transform, allows highly correlated regions to be compressed more efficiently than with the 4×4 transform. Most new videoconferencing products include H.264 (as well as H.263 and H.261 capabilities). Further extensions in version three of the original standard, 'Fidelity Range Extensions' (FRExt), were developed by the JVT (2005) to support higher-fidelity video coding through increased sample accuracy (including 10-bit and 12-bit coding) and higher-resolution colour information. The Fidelity Range Extensions project also included adaptive switching between 4×4 and 8×8 integer transforms, encoder-specified perceptual-based quantisation weighting matrices, efficient inter-picture lossless coding, support of additional colour spaces, and a residual colour transform (Marpe, Wiegand and Sullivan, 2006) to provide more options for efficient optimised designs. Perceptual-based quantisation scaling matrices defined in the FRExt take advantage of the HVS sensitivity to different types of error with the aim of improving subjective fidelity. Scalable Video Coding (SVC) was introduced as an amendment to the H.264/AVC Standard (ITU-T Recommendation H.264 & ISO/IEC 14496-10, 2005). SVC enables the creation of a video bitstream structured in layers. The SVC design enables decoding of the full bitstream or sub-sets created by removal of the enhancement layers. Scalability modes enabled by enhancement layer information include temporal, spatial and fidelity scalability (Schierl, Gänger, Hellge, Wiegand and Stockhammer, 2006).

A comparison of the performance of MPEG-2, H.263, MPEG-4 and H.264/AVC standards by Wiegand, Schwarz, Joch, Kossentini and Sullivan (2003) based on PSNR (Peak Signal to Noise Ratio, described in Section 5.1.2) and "informal subjective testing" demonstrated some important gains in coding efficiency in later versions of standards (and also the need for standards in subjective quality assessment methods, discussed further in Chapter 5). Tests in video conferencing applications showed that the H.264/AVC standard outperformed all of the other standards by a substantial margin with bit rate savings of more than 40% relative to H.263 (baseline) for video conferencing in PSNR tests. The "informal" subjective test results obtained by Wiegand *et al* (2003) found that sequences coded at 512 kbps with H.264/AVC were "subjectively equivalent" to the same sequences coded at 1024 kbps with MPEG-4 Visual. This represented a fifty percent bit rate saving which was larger than the savings indicated in the PSNR results. Subjective comparisons of pairs of videos (with

almost identical PSNR values) encoded at different bit rates using the different standards (MPEG-2, H.263, MPEG-4 and H.264/AVC) found that subjects indicated a significant preference for the H.264/AVC sequences and that this preference was highest for low bit rate encoded sequences. Although their subjective testing methods and results were not described in detail in the paper, Wiegand *et al* (2003) demonstrated the advantages of the latest standards in providing increased flexibility for optimising coding efficiency and ultimately improving user satisfaction. Although in the early stages of development, further perceptual benefits of FREXt have been described by Wedi and Kashiwagi (2004) in subjective quality testing of High Definition video content. In their experiments, Mean Opinion Scores (MOS, discussed further in Chapter 5) showed nominally better quality at a third of the bit rate (8 Mbps compared with 24 Mbps) and subjects reported that quality was difficult to distinguish from the uncompressed source video at 16 Mbps.

Video coding standards specify only the bit stream syntax and decoding process to enable interoperability. The design of the encoding process and the setting of coding parameters (for example, macroblock modes, motion vectors and transform co-efficient levels) are under the control of the developer (Wiegand, Schwarz, Joch, Kossentini and Sullivan, 2003). This provides the opportunity for flexible implementation and coding optimisation based on the application and quality of service needs of the user.

Lossy compression is essential for efficient transmission and storage of digital video data. In many applications, significant compression ratios can be achieved without loss of perceived quality but this is not the case for critical applications such as medical imaging (Raff, 1997) and task-specific applications such as BSL video communication where the quality requirements are such that loss can not be tolerated. The following section of this thesis reviews research and development of optimised solutions for achieving the perception of high quality video communication in lossy video coding systems.

4.2 Optimised Video Compression

Optimised video compression has been principally based on segmentation and coding prioritisation of a designated region (or regions) of ‘interest’ or ‘importance’ in the video scene using object or model based coding schemes.

Object-based coding is a feature of the MPEG-4 (ISO/IEC, 1998) standard (Section 4.1.4) which provides content-based functionality for the processing and compression of a video scene so that 'important' content can be detected and prioritised in the coding scheme. Such techniques have included segmentation based on shape, contour, colour (for example, skin colour) and motion. The segmented regions are given priority in the coding scheme resulting in higher compression ratios than uniform coding of the video images. The aim of these techniques is to exploit the limitations of the HVS to provide perception of high quality at lower bit rates (Sadka, 2002). For example, Soryani and Clarke (1992) applied a region-contour segmentation-based approach to locate the face for more accurate coding of this region in low bit rate video telephone applications.

Model-based coding uses a pre-defined model (for example, to represent a human face) which is adapted to detect objects in the video scene (Sadka, 2002). The model is adapted to match the contours of the detected object and the changes are coded to represent the object boundaries. This approach also has the potential to yield high compression ratios while maintaining good perceived quality but can only be used for sequences where the foreground object closely matches the predefined reference model (Choi and Takebe, 1994). Model-based encoding has been applied to the location, detection and tracking of faces in video communication systems (Eleftheriadis and Jacquin, 1995).

In sign language videoconferencing applications, ROI (Region-of-Interest) priority coding has been applied to optimise communications for deaf users at low bit rates. Schumeyer, Heredia, and Barner (1997) applied techniques to segment and give priority to a foreground region which included the 'sign space' (including the face and hands of the signer), in a 'static ROI' and by tracking the face and hands in a 'dynamic ROI', in a QCIF video sequence (this frame format was subsequently found not to be appropriate for sign language video, Section 3.3). Saxe and Foulds (2002) applied skin colour detection algorithms to segment sign language video frames into 'skin' regions (to include the face and hands) which were subject to 'moderate' compression and 'non skin' regions which were subject to higher degrees of compression. These examples assumed that the face and hands should be given equally high priority in the coding scheme.

Geisler and Perry (1998) demonstrated the potential to exploit the spatial resolution (particularly the foveal response) of the Human Visual System (HVS), in the implementation of an optimised video coding scheme (discussed in Section 2.3.2) while recognising that this required knowledge of location of gaze of the viewer. Foveated image and video coding techniques are closely related to ROI coding since an area of interest is defined around the point of fixation. Significant bandwidth savings can be achieved by this method since considerable high frequency information redundancy can be removed from the peripheral regions of the image without significant loss in perceptual quality of the reconstructed video image. Examples include Foveation Scalable Video Coding (FSVC) for wavelet-based compression (Wang, Lu and Bovik, 2003) and foveation applied in the DCT domain for block-based coding (Sheikh, Evans and Bovik, 2003).

Sheikh, Evans and Bovik (2003) proposed foveated processing techniques which employed ROI coding using an approximate model for foveation at the pre-processing stage and also in the DCT domain. They applied a "pre-processing engine" to calculate foveation regions and then applied a low pass filter to each region principally as a means of comparing the output with that of their DCT domain foveation method. The DCT foveation approach of Sheik *et al* (2003) used DCT sub-band weighting to suppress high frequency components in the peripheral regions of the image as determined by their approximate foveation model. They reported a reduction in complexity and bit rate in the DCT application but with more blocking artefacts and less suppression of higher frequency components than would be achieved by spatial image foveation.

Researchers at Bristol University, in collaboration with the University of Rochester in the USA, (Agrafiotis, Canagarajah, Bull and Dye, 2003; Agrafiotis, Canagarajah, Bull, Kyle, Seers and Dye, 2004 and 2006) developed methods of foveated and perceptually optimised video coding for sign language communication. They developed an approximate model of foveation by reducing the quantiser step size in the macroblocks in the face region and increasing the step size further away from the face, resulting in higher compression of the peripheral regions of the video image. A limitation of this approach is the problem associated with artefacts at block boundaries in a scheme where neighbouring macroblocks are encoded at different QP values. Their approach

requires rate control (due to the implementation of variable QP) to allow comparison of the performance of the method with standard CODECs at fixed bit rates.

4.3 Summary

Digital video data need to be compressed before transmission to optimise the bandwidth requirements of video communication systems. Video compression is the process of reducing digital video data to a smaller number of bits for storage and transmission. Lossy compression methods are applied to give the required compression ratios for video communication systems. Lossy compression removes subjective redundancy (that is, data which can be removed without significant impact on the viewer's perception of visual quality) and so the output reconstructed image is not identical to the source data. Removal of subjective redundancy is based on knowledge of the sensitivities of the HVS. This achieves high compression rates at the expense of overall video image quality. International standards in video coding have developed to improve efficiency and provide options for developers to optimise video communication systems to meet user demand for more services and better quality. In many applications, significant compression ratios can be achieved in standard systems without loss of perceived quality but this is not the case for task-specific applications such as BSL video communication where the quality requirements are such that loss interferes with the visual communication of information critical to the understanding of the language. Optimised video coding systems have been based principally on segmentation of a designated region of 'interest' or 'importance' in the video scene using objects or model based coding schemes. In sign language videoconferencing applications, ROI (Region-of-Interest) priority coding has been applied to optimise communications for deaf users at low bit rates. These methods have made assumptions about the visually important regions for BSL communication. Other approaches which exploit the foveal response of the HVS have been proposed but these have tended to adopt approximate models of foveation and the outputs are affected by artefacts at block boundaries.

The design of an optimised video communication system requires a trade-off between quality and the data rate achieved by compression. The measurement of video image quality is important for this trade-off decision and of critical importance to a system

designed for the specific communication needs of BSL users. The following chapter evaluates video quality assessment methods and defines the criteria for the design and implementation of subjective quality assessment in the experimental work in this thesis.

Chapter 5: Video Quality Assessment

Digital video images are subject to distortion and degradation during capture, processing, compression, storage and transmission resulting in reduced visual quality of the output of a video communication system. The performance of a digital video communication system also depends on the content of the input video signal (that is, the degree of motion and level of spatial detail in the image). The measurement of video quality is an active area of research which aims to determine the most robust and efficient methods of assessment. A suitable measure of quality was essential for optimisation of BSL video communication for sign language users. This chapter describes and evaluates objective and subjective methods of assessing video quality and establishes important criteria for the design of appropriate methods for BSL video quality assessment.

5.1 Objective Measurement of Video Quality

The aim of objective image quality assessment research is to develop quantitative methods which measure quality automatically. Objective image quality metrics are widely used in image processing for monitoring, benchmarking and optimising image quality (Wang, Bovik, Sheikh and Simoncelli, 2004). Image quality provided on networked communication systems can be monitored and adjusted by controlling and allocating resources. Video coding algorithms and parameters can be optimised using objective quality metrics; for example, in the optimal design of pre-filtering and bit allocation at the encoder and post-filtering and reconstruction at the decoder.

Objective evaluation techniques are mathematical models, designed to emulate the perceived quality afforded by the HVS, based on criteria and metrics that can be measured objectively. The objective methods are classified, according to the availability of the original video signal, which is assumed to be of high quality. There are three major classifications: full reference methods (where a complete reference image is assumed to be known); reduced reference methods (where the reference image is partially available) and no-reference methods (where no reference image is available).

5.1.1 Mean Squared Error (MSE)

The Mean Squared Error (MSE) method is the simplest full-reference quality metric applied in image processing. The MSE is computed by obtaining the mean of the squared intensity differences of distorted and reference image pixels (Equation 5.1). This approach considers an image signal as the sum of an original reference signal (I) and an error signal (K) and it therefore assumes that loss of perceptual quality is directly related to the visibility/strength of the error.

$$MSE = \frac{1}{mn}\sum_{i=0}^{m-1}\sum_{j=0}^{n-1}(I(i,j) - K(i,j))^2 \qquad \textit{Equation 5.1}$$

The calculation of MSE in Equation 5.1 is based on a monochrome image, where i and j denote the position in the two m×n arrays of pixels. For colour images with three RGB[11] values per pixel, the MSE is the sum of all squared value differences divided by image size and by three.

5.1.2 Peak Signal-to-Noise Ratio (PSNR)

The most widely applied full-reference objective method of evaluating the quality of digital video images is automatic computation of the Peak Signal-to-Noise Ratio (PSNR). It provides a fast indication of quality, based on error sensitivity and is used as a measure of the quality of reconstructed images in video CODECs (Richardson, 2002). PSNR measures the ratio between the maximum possible power of a signal and the power of the noise that affects the fidelity of its representation. PSNR is defined, with reference to the Mean Squared Error, in Equation 5.2.

$$PSNR(dB) = 10 \cdot \log_{10}\left(\frac{MAX_I^2}{MSE}\right) \qquad \textit{Equation 5.2}$$

[11] RGB - The RGB colour model is an additive model in which red, green and blue are combined to reproduce other colours.

In Equation 5.2, PSNR is measured in decibels (dB) and MAX_I is the maximum pixel value of the image. When the pixels are represented using 8 bits per sample, MAX_I is 255. More generally, when samples are represented using linear Pulse Code Modulation (PCM)[12] with B bits per sample, MAX_I is 2^B-1.

PSNR is normally expressed in terms of a logarithmic decibel scale (as signals tend to have a wide dynamic range[13] and the log scale provides a more realistic mapping to quality variations). The overall PSNR value of a video sequence is computed as the average PSNR for all the frames in the sequence. Generally, a high PSNR value indicates high video image quality and a low PSNR value is characteristic of low quality. Typical values of PSNR in image compression are between 30 and 40 dB and are normally reported to two decimal places. Comparison between two PSNR values gives an indication of the measure of quality improvement. Demonstration of a gain (dB) is used to indicate the value of an optimisation technique at a given threshold value which is considered to indicate an improvement.

5.1.3 Limitations of Automatic 'Objective' Measures of Video Quality

The MSE and PSNR methods measure quality in terms of error-sensitivity. This approach is objective since it is related to pixel luminance and chrominance values of the input and output video images (Sadka, 2002). These methods provide a measure of overall image quality in terms of error perception, but have a number of limitations.

Objective methods are applied after the encoding stage of video processing. This stage in the process may require several iterations in order to determine the encoding parameters that satisfy a specific level of user satisfaction, making them time consuming, complex and impractical for implementation in real-time applications. Winkler (2005) describes alternative objective evaluation approaches which enable the

---

[12] Pulse-Code Modulation (PCM) - a digital representation of an analogue signal where the magnitude of the signal is sampled regularly at uniform intervals and quantised to a series of symbols in a digital code. It is applied in digital telephone systems.

[13] Dynamic range - the ratio between the smallest and largest possible values of a changeable quantity.

prediction of the perceived quality level of an encoded video to be determined before the encoding stage to speed up the process for real-time applications.

For applications which do not require real-time monitoring, including the optimisation of algorithms and parameters in this thesis, objective methods do provide advantages in terms of speed, cost and consistency. However, objective quality methods (in general) have limitations as a reliable source of information about video image quality since they do not include subjective human intervention or accurate modelling of the complete human visual response to complex stimuli. Wang, Bovik, Sheikh and Simoncelli (2004) recognised the value of objective methods in image processing, and the limitations of the error-sensitivity approach for measuring the perception of quality. They devised an alternative method using structural-similarity based on the assumption that human visual perception is highly adapted for extracting structural information from a scene. Other approaches have included the development of a weighted-PSNR metric (Lee, Kwon, Jeong, Cho and Kim, 2002). However, these methods do not consider the human psychophysical and physiological factors which affect the perception of quality. Previous research (for example, Girod, 1992; Richardson, 2002; Zhong, Richardson, Sahraie and McGeorge, 2003) has demonstrated that PSNR does not correlate with measures of subjective video quality. Silverstein and Farrell (1996) concluded that there is only moderate correlation between perceived image quality and image fidelity (the ability of a process to render an image accurately), which casts some doubt on the objective definition of image quality for video applications and how it should be measured. They observed that an image may be 'enhanced' by a distortion. Subjects may detect the difference between an 'original' and its distorted version and, despite it having a lower PSNR, prefer the distorted version.

The foveal response of the HVS (Section 2.1.1) makes quality assessment much more complex than the simple linear model of behaviour which is the basis of PSNR calculations. Modelling the known characteristics of the HVS adds a significant burden in terms of computational complexity and is based on assumptions and generalisations which affect its use for coding optimisation based on user satisfaction. Objective metrics, such as PSNR, do not take account of the temporal visual response (Section 2.3.3) and so the impact of effects, such as frame rate and 'jitter', is not considered as a factor affecting overall image quality.

The main argument against the use of objective quality metrics, such as PSNR, for this thesis is that they are not able to account for the cognitive understanding and interactive visual processing (via eye movements) which influence the perceived quality of video images for a specific visual task such as sign language communication. It is argued that this can only be obtained using subjective quality methods.

5.2 Subjective Measurement of Video Quality

Subjective information obtained from users, experts or observers can be used as a measure of acceptable quality. This information is essential for the design of a video communication system for BSL users. An optimised video communication system, which aims to meet the needs of deaf users, must be inclusive by involving deaf subjects and communicating with them in the local version of their first/preferred language (BSL). However, according to Wang, Bovik, Sheikh and Simoncelli (2004), subjective quality evaluation is "inconvenient, time-consuming and expensive". Previous researchers have based their conclusions about quality on "informal subjective visual tests" (Wiegand, Schwarz, Joch, Kossentini and Sullivan, 2003), highlighting the need for rigorous methods which can be applied within the constraints of the research. Subjective methods may be structured or unstructured (Mullin, Jackson, Henderson, Smallwood, Sasse, Watson and Wilson, 2002). Structured subjective methods use formal data gathering instruments such as questionnaires, checklists and rating scales to efficiently gather concise data from users. Unstructured methods (including interviews, observation and post-hoc comments) produce free response answers. The advantage of these methods is that they provide wealth and depth of information. However, this approach requires extensive data analysis and the careful design of a research instrument which addresses specific design criteria (Table 5.1).

Subjective quality assessment may be considered in terms of measurement of different human responses (Mullin *et al*, 2002). These include measurement of task performance (for example, number of outputs, time taken or number of errors), user satisfaction (for example, measuring video quality against rating scales) and/or user cost (for example, physiological measurements of user states such as stress).

1	Validity – ability to measure exactly what is intended
2	Reliability and Repeatability – ability to provide consistent and repeatable results over a number of applications
3	Sensitivity – ability to measure the smallest variations in response as appropriate to the test
4	User Cost – ability to measure user response without causing disturbance to the user or interference with the task
5	Research Cost – ability to be applied within available resource constraints
6	Acceptability – ability to be acceptable as a means of gathering data by the user
7	Ease of Use/Application – ability to be applied within the expertise of the user

Table 5.1 Design Criteria for Subjective Assessment Methods (Adapted from Mullin et al, 2002)

Standard research instruments for the measurement of user satisfaction, using ratings scales, are described in Section 5.2.1 and evaluated in Section 5.2.2. Alternative methods (including measurement of task performance and user cost) are discussed in Section 5.2.3. A summary of methodological issues and the approach to obtaining subjective quality feedback from BSL users in this thesis is given in Section 5.2.4.

5.2.1 Standard Methods of Subjective Video Quality Assessment

The International Telecommunications Union (ITU), based in Geneva, is an international organisation which coordinates global telecommunication networks and services. It is the leading publisher of telecommunication technology, regulatory and standard information. The ITU currently has over 3,000 Recommendations and Standards (International Telecommunications Union, 2006) which define how telecommunication networks operate and inter-work.

The ITU Telecommunication Standardisation Sector (ITU-T) is one of the three Sectors of the International Telecommunication Union (ITU). The aim of the ITU-T Sector is to

ensure efficient and timely production of high quality standards (Recommendations) covering all aspects of telecommunications services.

The ITU-T Recommendations are not mandatory but compliance is widespread due to their high quality and because they are essential for the interconnectivity of networks and provision of telecommunication services on a worldwide scale. The ITU-T has two standards, currently in force, which address the subjective quality assessment of non-interactive, video applications.

ITU-T Recommendation BT.500-11 (2002) defines the methodology for subjective quality assessment of television pictures. The methods described include the Single Stimulus Method (SSM) and Double Stimulus Method (DSM) specifically for television viewing.

ITU-T Recommendation P.910 (1999) is more relevant to the specific nature of this thesis. It describes the recommended subjective video quality assessment methods for multimedia applications. The methods described in Recommendation P.910 are similar to those described for television viewing but are designed for specific applications, such as video conferencing, for a range of purposes including selection of algorithms, ranking audio-visual system performance and quality evaluation during an audiovisual connection. The recommendation specifies the standards for the source signal, video recording, sample size, experimental design, test methods, evaluation procedure and the analysis and reporting of results. This includes some specific recommendations about the length of video sequences (approximately 10 seconds) and the choice of test scenes (which should include the full range of temporal and spatial information of interest to the users of the device under test).

The recommended number of participants in subjective quality testing experiments is between four (the minimum for statistical reasons) and forty. The standards also provide a comparison of methods to enable the researcher to choose the most appropriate method for the requirements of the experiment depending on the context, purpose and point in the development process at which the test is conducted. Regardless of the selected method, the standards recommend that a short ‘mock test’ is conducted to familiarise the subject with the procedure before results are recorded. This is described as the ‘training phase’ and any subject feedback from this phase

should be considered in the design of the experiment. Three methods are described in Recommendation P.910.

5.2.1.1 Absolute Category Rating (ACR)

The Absolute Category Rating (ACR) method (or equivalent Single Stimulus Method as described in Recommendation BT.500-11) involves presenting test sequences one-at-a-time for subject rating of each individual sequence independently (that is, without reference to the original source sequence) on a five-level category scale. The scale for rating overall quality is given in Table 5.2.

Level	Rating
5	Excellent
4	Good
3	Fair
2	Poor
1	Bad

Table 5.2 Absolute Category Rating (ACR) Scale for Subjective Quality Assessment (ITU-T, 1999)

The Recommendation also provides additional options for a nine-level scale and for rating of quality other than for overall quality (such as visual object quality).

The test sequences are displayed to the user in sequence separated by a plain grey-coloured image which is displayed for a ten second period. During this period, the viewer must evaluate the test sequence shown using the ACR scale (Figure 5.1). Subjects are requested to view the entire video sequence before voting.

The experiment should be designed so that each test sequence is shown to the viewer for a specified number of repetitions (at least two further presentations of the same sequence are recommended) by repeating the same test conditions at different points in the experiment to ensure reliability and consistency of subject response.

The ACR method is described as being fast and easy to implement and well suited for qualification tests.

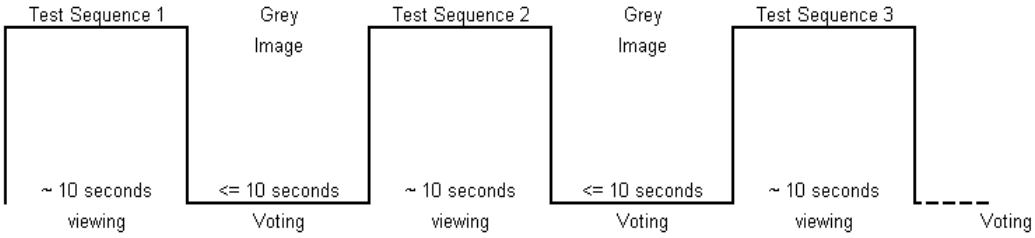


Figure 5.1 Time Pattern for ACR Stimulus Display (Adapted from ITU-T, 1999)

5.2.1.2 Degradation Category Rating (DCR)

The Degradation Category Rating (DCR) method (or Double Stimulus Impairment Scale Method) involves presenting test sequences in pairs; the first sequence in the pair is always the original (source) reference and the second sequence is created from the source and is the object of the test. The subject is asked to rate the difference in quality of the second sequence compared with the first sequence in the pair using a five-level category scale given in Table 5.3.

Level	Rating
5	Imperceptible
4	Perceptible but not annoying
3	Slightly annoying
2	Annoying
1	Very annoying

Table 5.3 Degradation Category Rating (DCR) Scale for Subjective Quality Assessment (ITU-T, 1999)

The test sequences in the pair are displayed to the user. There are two options for pair display depending on the frame format. In the case of full-screen images, video sequences should be presented to the viewer one-at-a-time separated by a plain grey-coloured display image which is displayed for a two second period. In the case of reduced frame formats (such as CIF or QCIF), the video sequences should be displayed simultaneously on the same monitor with a plain grey background. Each pair of video sequences should be separated by a plain grey-coloured display image which

is displayed for a ten second period. During this period between pairs, the viewer must compare the displayed test sequences using the DCR scale (Figure 5.2).

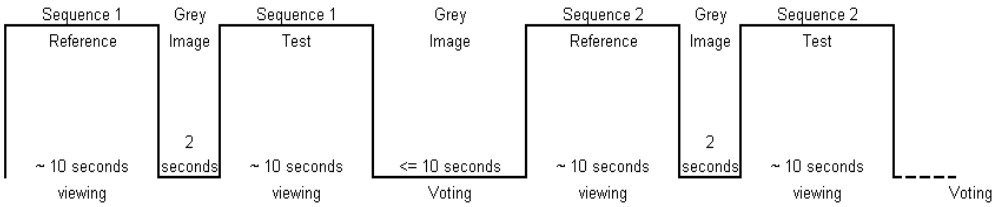


Figure 5.2 Time Pattern for DCR Stimulus Display (Adapted from ITU-T, 1999)

The experiment should be designed so that each test sequence is shown to the viewer for a specified number of repetitions (at least two further presentations of the same sequence are recommended) by repeating the same test conditions at different points in the experiment to ensure reliability and consistency of subject response.

The DCR method is recommended for testing fidelity of transmission with respect to the source signal for situations where the viewer's detection of impairment is of importance.

5.2.1.3 Pair Comparison (PC)

The Pair Comparison (PC) method involves presenting test sequences in pairs. The pairs are created from the same source presented through two different systems under test. Unlike the DCR method, the order of display of sequences in the pair is not important (since there is no source reference) but the experiment should be designed so that all pairs of sequences are displayed in both possible orders. This procedure means that the necessary repetitions are built in to the design of the experiment and so, unlike ACR and DCR, further replications of experimental conditions do not need to be considered.

The test sequences in the pair are displayed to the user. As with the DCR method, there are two options for pair display depending on the frame format. In the case of full-screen images, video sequences should be presented to the viewer one-at-a-time separated by a plain grey-coloured display image which is displayed for a two second period. In the case of reduced frame formats (such as CIF or QCIF), the video sequences should be displayed simultaneously on the same monitor with a plain grey

background. Each pair of video sequences should be separated by a plain grey-coloured display image which is displayed for a ten second period. During this ten second period between pairs, the viewer must express preference for the quality of either Sequence 1 or Sequence 2 in the pair (Figure 5.3).

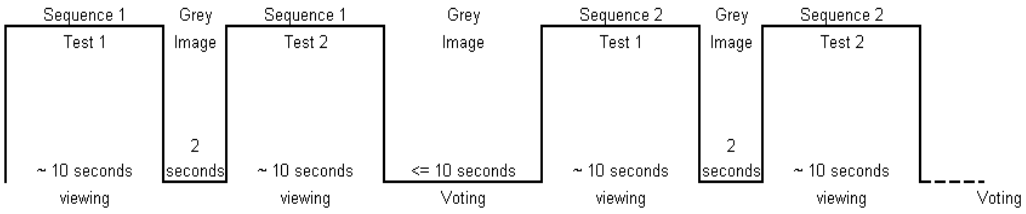


Figure 5.3 Time Pattern for PC Stimulus Display (Adapted from ITU-T, 1999)

The PC method is recommended for discriminating between test items which are similar in quality. This method is not recommended for a large number of test items due to the time needed to conduct the test with each individual subject. The standard recommends that in this case an ACR or DCR test should be conducted followed by a PC test on those items which received a similar rating.

The subjective quality testing methods described in this section provide a standard approach to obtaining qualitative feedback from users. The advantage of this approach is that the quality scales are easy to administer and score (Mullin *et al*, 2002). The following section evaluates the limitations of these methods.

5.2.2 Limitations of ITU Methods for Subjective Video Quality Assessment

The ITU recommended scales are primarily designed to determine whether subjects can detect degradation in picture quality rather than with respect to user task. The limitations of the ITU recommended scales for measuring subjective quality assessment of real-time interactive communication applications were identified by Watson (2001) and reported by Mullin *et al* (2002). The criticisms of the ITU methods, relevant to this thesis, concern the duration of the test sequences and the rating scale descriptors. They argued that the recommended ten second test video clip was not of sufficient duration for assessing video quality for real-time Internet video where there is

likely to be a large variety and range of rapidly changing impairments. However, one of the main weaknesses of subjective methods is the time taken to conduct the test and so short clips are preferable from the perspective of overall experiment duration, particularly where there are a number of conditions to be tested in controlled conditions. Aldridge, Davidoff, Ghanbari, Hands and Pearson (1995) found that assessors gave more 'weight' to the content near the end of the sequence in their quality assessment using DSCQS (Double Stimulus Continuous Quality Scale) tests for television pictures. This observation gives further support to the argument for keeping test sequences short. The test sequences used in the experiments conducted for this thesis were created to ensure that each clip conveyed meaningful information in naturally expressed BSL in a period of approximately ten seconds. This kept the overall time for the experiments to a minimum and maintained the interest of the subject in the video material.

The vocabulary used in the ITU scale point descriptors (Tables 5.2 and 5.3) is expressed in general terms (for judgement of impairments) and is not relevant to a specific video communication task. Researchers have also found that subjects find it difficult to identify the exact point on the scale which reflects their opinion. In DCR tests, it may be the case that a difference between the original and test sequences is 'perceptible but not annoying' (equivalent to a rating score of 4, Table 5.3) when the user may actually perceive the difference but prefer the test sequence to the original (that is, better than a rating score of 5). In ACR tests, Jones and McManus (1986) observed that subjects found it difficult to distinguish between 'fair' and 'good' ratings and Vertanen, Gleiss and Goldstein (1995) observed difficulty in distinguishing between 'bad' and 'poor' ratings. This effect is pronounced in multi-national subject groups where interpretation of the translated descriptors varies considerably (Jones and McManus, 1986). It is also the case that there are no distinct signs for each of the ITU rating scale descriptors in BSL.

Jones and McManus (1986) argued that the intervals between the ITU scale point ratings are not equal in size. This was confirmed by Watson (2001) and Mullin *et al* (2002) who demonstrated that the ITU scale descriptors do not represent equal perceptual intervals. This evidence implies that parametric statistical tests (for interval data) could not be legitimately performed on the data gathered from these tests.

The ITU scales provide a standard method of assessing video quality and a framework for the design and execution of experiments with end-users in a controlled experimental environment. This approach has limitations for some applications, particularly real-time Internet video and for specific tasks such as BSL video communication. The guidelines in the ITU Recommendation P.910 acknowledge the difficulties with translation of the rating scale descriptors in different languages and recommend the use of continuous scales as an alternative to overcome the language barrier. The 'Quasi-Continuous' scale described in the standard uses a numerical scale with labels at the start and end points of the scale, and an unlabelled marker at the mid-point of the scale. However, this does not ease decision making for the assessor and does not allow evaluation of quality in relation to a specific task.

The importance of developing suitable subjective (and objective) methods of assessing video image quality for the growing number of multimedia communication services is acknowledged by the ITU in their continuing work in this field. ITU-T Study Group 9 (Integrated Broadband Cable Networks and Television and Sound Transmission) was tasked with the development of new methods for evaluating perceptual quality of multimedia services (Question 14/9) with the aim of producing a new/updated recommendation by the end of the study period in 2008 (ITU-T Study Group 9, 2004).

Alternative methods for video quality assessment, including the application of continuous rating scales, are reviewed in the following section.

5.2.3 Alternative Methods of Video Quality Assessment

Continuous rating scales, such as Single Stimulus Continuous Quality Evaluation (SSCQE) methods (ITU-T Recommendation BT.500-11, 2002), may be applied to allow time–varying image quality perception feedback to be collected from subjects throughout a video session. In a continuous assessment experiment, the user views the video material and moves a slider up and down to record quality perceptions during the session. The position of the slider is recorded at regular time intervals during the experiment. Bouch, Watson and Sasse (1998) developed a QUASS (QUality ASsessment Slider) for continuous quality rating on a 0-100 scale (controlled by the user with a mouse) with results recorded every second during the experiment. The QUASS software tool was not based on ITU rating scales and was applied to directly

map objective and subjective quality metrics and overcome the limitations of the ITU rating scale descriptors (Section 5.2.2).

Richardson and Kannangara (2004) developed a fast User Feedback Quality (UFQ) method for obtaining subjective quality feedback measurements from users choosing between a number of alternative coding or processing options. The UFQ method enables the user to select 'best' quality parameters automatically by moving a slider until the preferred video clip quality is obtained. In this method, the slider position is recorded (sampled) at regular time intervals and the final slider position determines the test subject's preferred choice of parameters. The UFQ method is a fast and robust method for measuring visual quality where there is a choice between a wide range of coding parameters (such as different frame rates), alternative coding algorithms or alternative post-processing algorithms.

However, a general disadvantage of this type of approach is that operation of a slider may be a distraction during passive viewing or a barrier to the user performing a real visual task (McCarthy, Sasse and Miras, 2004) such as BSL communication in this thesis. McCarthy, Sasse and Miras (2004) identified the requirement for a method which was easy to apply, not disruptive to the user task and would "elicit continuous ratings of quality with minimal effort on the user's part". They used eye tracking (to determine regions of interest during the video stimulus) and a simple verbal binary acceptability rating (acceptable or unacceptable) or 'method of limits' (that is, a method which determines detection thresholds by gradually increasing/decreasing the intensity of a stimulus in discrete steps until it is distinguishable) for the duration of the experiment. McCarthy *et al* (2004) applied the method during a continuous stimulus where quality changed over time (every 30 seconds). Subjects were asked to identify the point at which data became 'acceptable'. The advantages of this approach are that it is easy for users to understand, it is less disruptive to users than other continuous assessment techniques (for example, using a slider) and it can be used with variable video quality which is relevant to service providers. The disadvantage of this approach is the precise determination of the rating awarded by the user at specific points in the video sequence.

Subjective methods are widely used to determine user satisfaction. Measurements of user satisfaction are cognitively mediated and are influenced by external factors such

as financial cost associated with quality and task difficulty (Mullin *et al*, 2002). The measurement of task performance is another option for obtaining subjective feedback from users. Dugénie, Munro and Barton (2002) considered a case study of video telephony for deaf people in their research on assessing subjective Quality of Service (QoS) of mobile multimedia applications delivered over the Internet. They based their subjective analysis on the level of comprehension of the transmitted sign language, rather than on the degree of satisfaction with the service (which was found to be influenced by technology assumptions and opinions about the manner in which the sign language was communicated). Dugénie *et al* (2002) concluded that their approach was scalable and therefore easier to analyse than subjective methods that produced large amounts of qualitative data. However, they found that there was a wide spread of results amongst the sample of ten subjects and they recommended that a larger sample be used to increase confidence in results of further experiments. It may also be the case that satisfactory communication for the individual user may not correlate with the level of comprehension of the transmitted sign language (measured by the number of correct responses to questions about the content of the sign language transmission) and so this method was not applied in this thesis.

An alternative approach (which may be applied separately or with measurement of user satisfaction and task performance) is to measure user cost. User cost is the level of stress and discomfort experienced by the user when presented with poor quality video during a visual task. User cost may be assessed subjectively using mood scales similar to the ITU rating scales. However, these present the same limitations as user satisfaction scales and, since they are also cognitively mediated, have limited value in addition to measures of satisfaction and task performance. Wilson and Sasse (2000) argued that autonomous physiological human responses, as *objective* measures of user cost, are not subject to cognitive mediation and may be applied without interfering with the primary visual task. They obtained physiological measurements (Heart Rate, Blood Volume Pulse and Galvanic Skin Resistance) as indicators of stress during media quality assessment tests which also employed the QUASS tool for continuous quality rating (Bouch, Watson and Sasse, 1998). They found that low frame rate video content produced a stress response which was not picked up in subjective tests of user satisfaction and concluded that quality assessment should employ a combination of methods, including physiological measurement of user cost. Wilson and Sasse (2000)

demonstrated that objective physiological methods were unobtrusive and that accurate indicators of user cost could be obtained based on baseline physiological readings taken prior to the experiment and by experimental design which carefully controlled the test environment. However, they acknowledged that it may be difficult to separate the stress caused by the test conditions from the emotions arising from the video content and that the sensors could interfere with interactive tasks (such as typing). In this thesis it is argued that physiological measures of user cost/stress would distract the subjects from the primary task and that they would be difficult to separate from other emotions arising from the BSL content in the test video sequences.

5.3 Summary

The aim of this chapter was to review available methods of video quality assessment and identify criteria for the design of a suitable method for end-user evaluation of video quality for BSL communication.

Automatic, 'objective' image quality metrics are widely used for optimisation of video communication systems but are limited since they do not accurately model the complete visual response to complex visual stimuli. They are quantitative measures of image fidelity rather than measures of perceived quality for a specific purpose or task.

Subjective video quality assessment methods provide more valuable information about perceptions of quality by end-users or observers. However, they are regarded as being time-consuming and the approach tends to be informal. They also require design of a suitable research instrument and extensive data analysis. Important design criteria for subjective quality assessment have been identified (Mullin *et al*, 2002) which consider the reliability and validity of the method and the demands on the user/observer. Standard (ITU) methods of subjective quality assessment specify the procedures and rating scales for detection of impairments and comparing and discriminating between systems. These methods enable the capture of qualitative feedback from users but are designed to assess the response to degradation in picture quality rather than quality for specific visual task. Problems have been identified with the timing of stimulus displays, the vocabulary used in the rating scale descriptors (including translating these into different languages) and the perceptual intervals between points on the rating scales. This has led to the development of alternative (non-standard) methods.

Non-standard subjective methods, including continuous quality rating using a slider and user acceptability tests using a binary 'method of limits', have been developed in previous research to address some of the problems of standard methods. Objective measures of task performance and physiological responses (measures of user cost/stress) have also been applied in previous research to address the limitations of current subjective methods for specific applications. However, most of the alternative methods reviewed were considered unsuitable for application in this thesis since they might distract the user from the primary visual task (BSL communication). The conclusion of this review is that a new approach is required which combines the rigour and reliability of procedures in standard ITU methods for subjective testing with alternative methods of capturing feedback from the user about perceived quality for the visual task. The specific design criteria and development of the methods applied in this thesis are described and evaluated in Chapter 9.

PART TWO: EXPERIMENTAL WORK

Chapter 6: Visual Response to BSL Video Image Content

Eye Movement Tracking (EMT) is used in a wide range of research areas, including Cognitive Neuroscience, Psychology, Industrial Engineering and Human Factors, Computer Science and Human-Computer Interaction (HCI). It is also used in advertising and marketing research and in medical research for a wide range of diagnostic and interactive purposes including analysing cognitive intent, interest (ROI) and salience (Duchowski, 2002). EMT to determine the visual scan path of viewers during a range of tasks is discussed in Section 2.2.3.

An optimised video communication system that meets the needs of deaf users must include a study of viewing behaviour during a sign language task. Consideration of characteristic viewing behaviour of deaf people and the study of eye movements, together with the visual attention mechanisms required for active BSL communication, are essential for identification of 'important' image content of BSL video for selective prioritisation in a video coding scheme. EMT is applied to determine how deaf people sample the visual scene during a BSL communication task and to investigate:

1. The visual scan path (sequence of saccades and fixations) and whether eye movements are characteristic and consistent for particular stimuli for (and between) individual subjects.
2. The visually important image content, viewed at high acuity at the point of gaze, and the image content viewed at low resolution in peripheral vision.
3. The visual response to the characteristics of the BSL stimulus content.

This chapter describes the eye tracking methods and methodological issues for eye movement data capture and analysis. The operational procedures for capture of eye gaze data and the techniques developed to display and analyse the captured data in this thesis are described for two EMT experiments. The first experiment (Section 6.3) was conducted to determine whether there was a typical visual response to sign language video content and to identify regions of 'importance' for the viewer. This research was presented at a seminar and international conferences (Muir and Richardson, 2002; Muir and Richardson, 2003 and Muir, Richardson and Leaper, 2003).

The second experiment (Section 6.4) produced more detailed analysis of EMT data with respect to image content and was published in an international journal (Muir and Richardson, 2005). Conclusions based on the results and analysis of the EMT experiments are synthesised in Section 6.5.

6.1 Eye Movement Tracking Methods

There are different types of eye tracker designed to suit different applications and budgets. These range from fixed-head systems (using a chinrest to keep the head steady), head-mounted systems (which allow some free head movement) to systems which function remotely and automatically track the head during motion (for example, in the system developed by Tobii (http://www.tobii.se/) the computer screen detects, captures and tracks eye gaze). The less intrusive, high cost remote systems are principally applied as input devices in computer games (Virtual Reality) but also in large scale medical, psychological and vision research projects, and marketing impact studies. Fixed-head and head-mounted systems provide a lower cost option for reliable tracking in science and humanities research projects. The most widely available and affordable devices are video-based corneal reflection eye trackers (Duchowski, 2003). In this type of system a camera focuses on one (or both) eyes and records eye movements while the viewer looks at a stimulus (still image, video sequence or PC user interface). Most modern eye-trackers use contrast to locate the centre of the pupil and use infrared light to create a corneal reflection. The vector between the pupil location and the corneal reflection (glint) is used to compute gaze intersection with a surface after calibration for each individual subject. Two general types of eye tracking techniques are Bright Pupil and Dark Pupil. The difference between these techniques is the location of the illumination source with respect to the optics. If the illumination is coaxial with the optical path then the retina of the eye reflects the light, creating a bright pupil effect similar to 'red eye'. If the illumination source is offset from the optical path, then the pupil appears dark. Bright Pupil tracking creates greater iris/pupil contrast for more robust eye tracking and reduces interference caused by eyelashes and other obscuring features. It also enables tracking in lighting conditions ranging from total darkness to very bright, although it is not effective for tracking outdoors (due to interference with extraneous IR sources). In Dark Pupil tracking methods, the pupil acts as an infrared sink so that it appears as a 'black hole'. This method is simpler to apply

and is less sensitive to head movements on the z-axis (that is, movement towards or away from the camera) (Arrington Research, 2002).

Technical specifications of eye tracking devices vary considerably depending on the system and the purpose for which they are being used. Eye tracking systems use a sampling rate of at least 30Hz, although 50/60 Hz is more common in devices designed to capture the detail of very rapid eye movements.

The eye tracking systems used in this thesis were the Eye Science™ Gaze Tracker (EyeTech Digital Systems, http://www.eyetechds.com/) and the ViewPoint™ Eye Tracker (Arrington Research Inc., http://www.arringtonresearch.com/).

The Eye Science™ system was provided on loan and was used in the first set of eye movement tracking experiments (Section 6.3). This is a monocular video-based system using the dark pupil tracking technique and infrared illumination. It was hosted on a PC running Windows 98 and the video camera and infrared lights were mounted on the PC monitor with nothing attached to the subject. The maximum sample rate was 30Hz. Small head movements were possible with this system but the subject's eye had to be maintained in the camera field of view (4×4 cm). To assist the subject to minimise head movements, a comfortable chair-mounted headrest was provided. The eye gaze data were saved to an ASCII file for further analysis. The Eye Science™ system was adequate for the initial eye tracking experiment but was replaced by the ViewPoint™ Eye Tracker purchased for the second set of experiments (Section 6.4).

The ViewPoint™ Eye Tracker (Arrington Research, 2002) used infrared video for monocular tracking by either Bright Pupil or Dark Pupil methods (set by the researcher). The sample rate was selectable between 30-60Hz. Head movements were restricted using a QuickClamp™ chinrest which could be adjusted to comfortably accommodate different face and head shapes and sizes (Figure 6.6). Small head movements were permitted in this system as long as the subject's pupil and corneal reflection remained within the camera image. Most manufacturers of eye tracking devices claim that they can tolerate some head movement. However, previous research demonstrated that more accurate tracking is obtained by constraining head movements within comfortable limits (Jacob, 1991). The technical specifications for the ViewPoint™ Eye Tracker are detailed in Appendix B. The eye gaze data including (x, y) coordinates of gaze, pupil

height and width, delta and total time were stored in ASCII format and input to Matlab software (http://www.mathworks.com/) for display and analysis.

6.2 Methodological Issues in Eye Movement Tracking

Significant advances in eye tracking methods in recent years have led to the development of systems which are less intrusive for subjects and more accurate (due to more sophisticated algorithms for determining eye positions and fixations). However, researchers have identified a number of technical limitations, operational issues and assumptions which need to be addressed in the design of eye tracking experiments (Spiller, 2004). Duchowski (2003) described the complexity of installation, set-up, calibration, operation, data collection and analysis of data in his review of the main types of eye tracking devices. Rayner (1998) identified the lack of standardisation as a potential issue for eye tracking measurement and methodology. He found significant variation in the way that eye movements were monitored and analysed, while acknowledging that there was evidence of repeatability of important findings across research projects. The identification of fixations is an important aspect of eye movement data analysis and is principally a subjective process. Salvucci and Goldberg (2000) proposed a taxonomy of fixation identification algorithms for research and development of future eye-based systems and applications. Jacob (1991) reported that researchers encounter problems with tolerance to head movements, over-sensitivity to noise (and other external factors) and loss of calibration. Problems with initial calibration for individual subjects were also reported by Hyőnä and Lorch (2004). These researchers used an infrared video-based tracking system mounted on a headband and reported that data for seven of the sixty-six participants were excluded due to calibration problems. Schnipke and Todd (2000) reported problems in collecting eye tracking data, despite vendor training and one year's experience, using a remote eye-gaze system. They identified problems with ease of use of the system and calibration. They obtained eye tracking data successfully from only six of sixteen participants (a 37.5% success rate).

Webb and Renshaw (2005) identified three key methodological issues for commercial use of eye tracking which were relevant to this thesis:

1. Cognition and visual attention – eye movement tracking forces an assumption about how foveal attention and cognitive processes are linked. The relationship between eye movements and attention was discussed in Section 2.2 of this thesis. Webb and Renshaw (2005) stated that it is generally assumed that attention is indicated by eye gaze when analysing eye movement data and that this has implications for analysis of eye tracking data. They argued that a long fixation may indicate a ROI but it may also indicate that the subject is 'puzzled by what the object means'.
2. Behaviour, emotion and visual attention – the human response to a visual stimulus includes physiological responses (for example, body temperature and heart rate) which give an indication of the physical cost to the stimulus viewer (Riegelsberger, Sasse and McCarthy, 2002). Physiological measurement techniques are being developed for eye tracking systems for consumer and usability research, for example, skin-conductivity sensors for measuring changes in perspiration, temperature and heart-rate (Eye-Square, http://www.eye-square.com/) and pupil diameter recording as a measure of emotional response (Eye Tracking Inc., http://www.eyetracking.com/).
3. Eye tracking metrics – a number of metrics can be used to analyse eye gaze data; total fixation time on a region of interest (for example in a study of ASL by Thompson and Emmory, 2003), the gaze path (for example in web-page scanning by Nielsen, 2006) or the duration of the first fixation (for example, in image impact analysis in commercial applications). The metric used depends on the nature of the task and the stimulus.

The eye tracking experiments conducted with deaf people viewing sign language video sequences in this thesis were designed to address the methodological issues raised by previous researchers. As there was no requirement for subject interaction during the experiments, comfortable head restraint was included to maximise the accuracy of calibration and data collection during the experiments. In the analysis of the eye movement data it was assumed that, for the complex task of BSL communication, there is a strong relationship between gaze location and attention (discussed in Section 2.2 of this thesis). Fixations (Section 2.2.2) were identified as being of at least 200ms in any one location (ROI). They were measured in terms of the percentage fixation time on regions of importance and analysed with respect to image content, optical flows

(Section 7.1) in the video image stimuli and user feedback (comments communicated in BSL after the eye tracking sessions). Physiological measurements were not measured in this thesis as it was considered that emotional response to image quality could not be separated from the emotions arising from the BSL video content.

Goldberg, Stimson, Lewenstein, Scott, and Wichansky (2002) identified two styles of eye tracking approach: top-down (task oriented) and bottom-up (behavioural inferences). They concluded that both of these approaches must be adopted for the use of eye-tracking as a usability methodology. The BSL communication task and the influences of the HVS and BSL communication behaviour were of key importance in the design of the experiments and analysis of the eye tracking data in this thesis.

6.3 EMT Investigation of Regions of Importance

This section describes the application of EMT to investigate the visual responses of profoundly deaf viewers during BSL video communication.

6.3.1 Experimental Design and Rationale

An EMT experiment was designed to record and analyse the gaze points of sign language users whilst viewing a BSL video sequence. The aim was to identify the visually important regions of sign language video images for BSL video communication. In this experiment, profoundly deaf adult participants, who used BSL as their first language, observed a sixty-second video sequence with the task of understanding the signed story conveyed by the signer in the video. The sign language video test sequence was played twice for each subject. The aim of this repetition was to test whether the subjects exhibited a consistent response to the same re-presented video stimulus (discussed in Section 2.2.3). Eye movement data were captured for each display of the sequence for each participant and analysed to identify regions-of-importance.

6.3.2 Method

EMT apparatus was used to capture the gaze points of deaf people viewing a sixty-second BSL video sequence.

6.3.2.1 Subjects

The experiment was conducted with nine profoundly deaf volunteers from the Aberdeen Deaf Social and Sports Club (ADSSC). For each subject, BSL was the preferred/first language and so all communications were in BSL, aided by a local qualified BSL Interpreter who was well known to the participants. Data for one of the volunteers was excluded from the final results due to loss of eye movement tracking during the experiment. The eight subjects who completed the experiment included four males and four females aged between thirty-six and sixty-three years.

6.3.2.2 Apparatus

An Eye Science^{TM} Quick Glance Gaze Tracker (EyeTech Digital Systems, http://www.eyetechds.com/) was used to track and record the eye movements of deaf subjects. The Eye Science^{TM} equipment is a monocular video-based system which applies dark pupil tracking technique and uses infrared illumination (described in Section 6.1). The eye tracker was hosted on a PC running Windows 98. The video camera and infrared lights were mounted on the PC monitor with nothing attached to the subject. The maximum sample rate was 30Hz. Small head movements were allowed but a comfortable chair-mounted headrest was provided to assist the subjects to restrict their head movements.

6.3.2.3 Materials

The test video sequence was a sixty-second clip recorded at 25 frames per second. The signer in the video described her experience as a child in the north of Scotland watching her mother speaking on the phone and her feelings of missing the pleasure of communicating with distant friends and family. The signer remained in the same position for most of the clip apart from stepping forward and looking down to turn a page of her notes for a short period. She used a wide range of BSL movements and gestures during the signed story. Figure 6.1 shows two frames from the test sequence which illustrate finger-spelling and use of body movement in the test video sequence.

Figure 6.1 Sample Frames from the Test BSL Video Sequence

6.3.2.4 Procedure

The infra-red sensitive camera, secured to the chassis of the monitor (directly below the screen), was focused and the infra red LEDs (Light Emission Diodes), mounted on each side of the computer monitor, were positioned until the eye was in clear focus and brightly illuminated by infra-red light. The eye tracker was calibrated, using sixteen on-screen visual targets, for each of the profoundly deaf volunteers before the experiment commenced. The test sequence was displayed full screen on a PC monitor and each participant watched the video from a comfortable viewing distance. The sequence was played twice for each subject (with a sequence of calibration markers between the two instances). The sequence had not been seen by any of the subjects before the experiment. The (x, y) gaze point coordinates of each subject were recorded by the EMT software and saved to a unique data file for analysis.

6.3.3 Results

The EMT data for each subject were plotted on the test video sequence frames (using Matlab software). The results showed a consistent response across each of the subjects and for each instance of the test video display. The set of (x, y) gaze coordinates for three of the test subjects (Subject A, Subject B and Subject C) are plotted in Figure 6.2.

The eye movements of the eight subjects exhibited significant similarities. In each case, the gaze was concentrated on the face of the signer in the video sequence, with

occasional visual ‘excursions’ to other regions mostly on the vertical axis through the face of the signer (see plots for Subject A and Subject C). Two of the subjects in the sample group (one of which was Subject B) showed very precise concentration on the face with very few excursions.

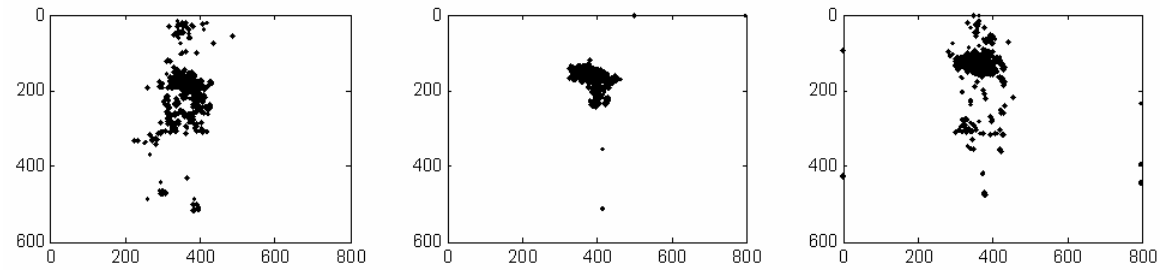


Figure 6.2 Scatter Plots of EMT Data for Subjects A, B and C

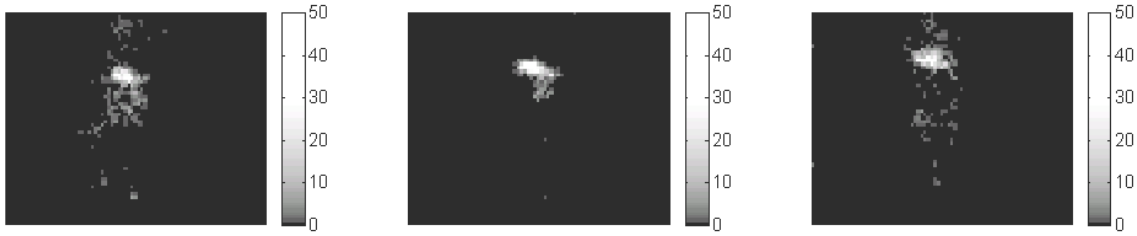


Figure 6.3 Histograms of EMT Data Densities for Subjects A, B and C

The results for Subjects A, B and C are presented as 2-D histograms in Figure 6.3, which plot the number of gaze points occurring in each 10×10 pixel square in the image. The majority of the points occur around the face region of the signer (‘bright’ area) with a small number of gaze points outside this region.

Figure 6.4 shows the scatter plot for Subject B overlaid with circles representing viewing angles relative to a central point. The circles are centred on the median position of all the samples and are plotted at constant angles of 2.5^{o}, 5^{o} and 10^{o} from the centre. It is clear from Figure 6.4 that most of the gaze points fall within 2.5^{o} of the centre (the angle of high resolution vision of the fovea): over 75% of points lie within this circle for Subjects A and C and over 90% for Subject B. This indicates that the viewer was getting a sharp view of the face and that other regions of the image were seen in peripheral vision until a rapid eye movement (saccade) moved the fovea away from the face region.

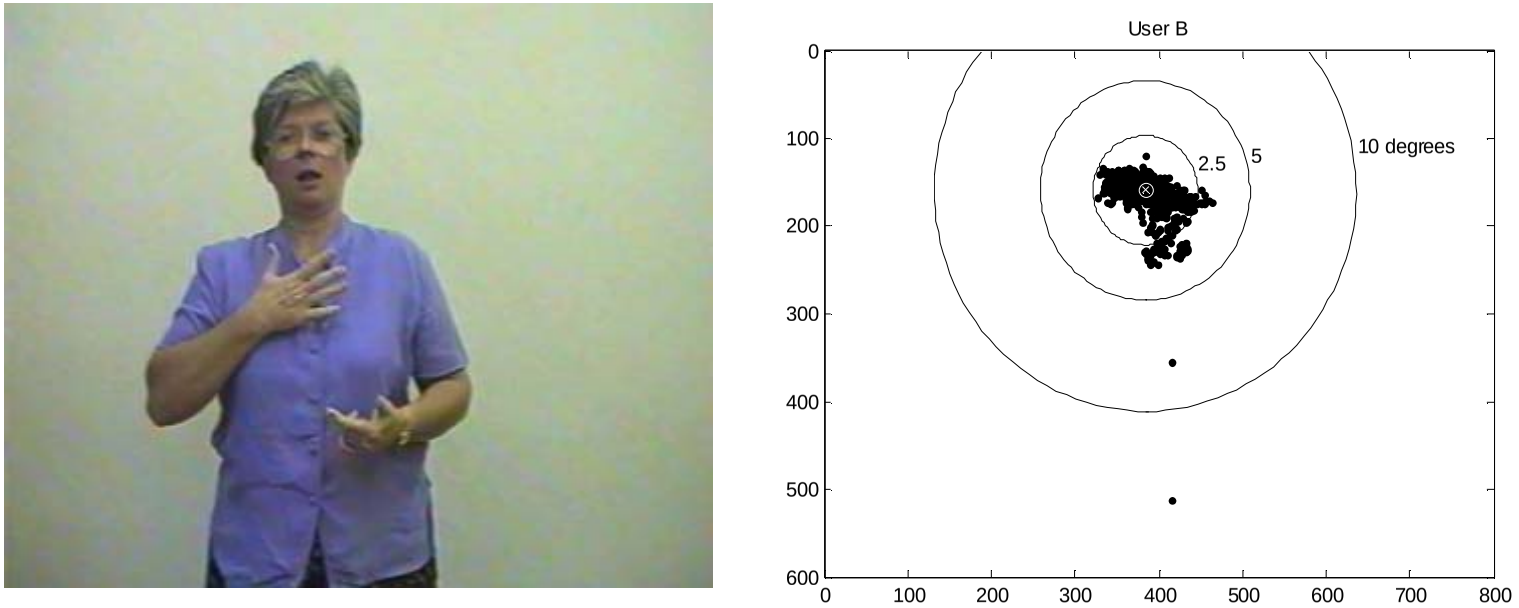


Figure 6.4 Sample Frame and Scatter Plot of EMT Data for Subject B Showing Angular Distribution

In order to investigate the vertical excursions (away from the face of the signer) observed in the data, the Y-coordinates of the subject data were plotted over time. Figure 6.5 plots the Y-coordinates of gaze points for Subject A against time for the test video sequence.

The median Y-coordinate is 185 for Subject A and the plot clearly shows that gaze is concentrated mostly around this position (on the face of the signer) with occasional visual excursions. Excursions of the subject's gaze away from the median typically last for less than 0.5 seconds and are concentrated in the vertical axis through the centre point. At one point in the video sequence no signing takes place when the signer looks down and turns a page of her notes. This corresponds to a large downward visual excursion by Subject A (at around 95 seconds on the time axis in Figure 6.5). There is a similar significant excursion at the same point in the sets of results for the other seven subjects. The remaining excursions are less significant. For the very short excursions (saccades), it is most likely that the visual stimulus was suppressed during this time (discussed in Section 2.2.1). The very small number of visual excursions, longer than 0.5 seconds in duration, tended to fall below the median point, in the region of the lower face or just below the face of the signer in the video.

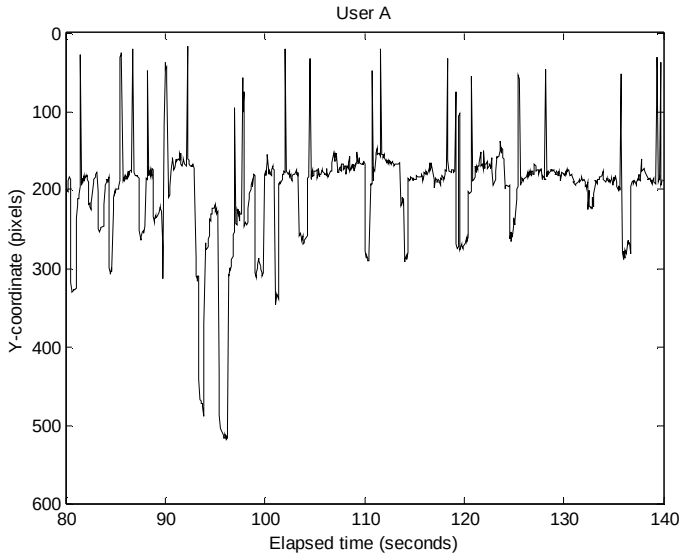


Figure 6.5 EMT Data (Y-Coordinates) over Elapsed Time (seconds) for Subject A

6.3.4 Discussion

The EMT results indicated a response to the visual stimulus that was consistent across all subjects. In each case the gaze location corresponded mainly with the position of the face of the signer. This supports the fact that facial expression, lip/mouth shape and eye position convey important information during a sign language conversation and these features require more detailed (foveal) resolution than the more expansive movements of hands. Most of the recorded gaze points occurred within a viewing angle of 2.5^{o} relative to the centre point. The effective resolution of human vision drops by a factor of two at a viewing angle of 2.5^{o} from the point of attention and reduces logarithmically beyond this angle. Five of the subjects' responses included some excursions from this central area, usually along the vertical axis, lasting up to 0.5 seconds but typically around 150-200ms. During a short eye movement (or saccade), high-resolution vision is suppressed by the Human Visual System (saccadic suppression, discussed in Section 2.2.1) and so the image is not perceived by the viewer during that time.

These results implied that an experienced sign language user perceives the face region with high visual resolution throughout a sign language conversation in order to extract information from the face, lips/mouth and eyes. Hand and body movements are perceived mainly in lower-resolution peripheral vision. Occasional saccadic excursions away from the face area were made (by some subjects). However, as vision is suppressed during saccades, the hand/body region was not observed in full spatial detail.

The results of this experiment demonstrated that experienced BSL users exhibit a consistent, characteristic eye movement response to BSL video and that the face is the most visually important region of the stimulus. More precise location and analysis of eye gaze was an objective for the second EMT Experiment.

6.4 EMT Investigation of Visual Responses to BSL Content

This section describes the application of an EMT experiment to investigate the visual responses of profoundly deaf subjects to BSL video image content.

6.4.1 Experimental Design and Rationale

A second EMT experiment was designed to build on the first EMT experiment (Section 6.3) by extending the study to include more participants, a wider range of sign language video material, more detailed analysis of the regions of 'importance' and identification of factors that influence the visual attention of deaf people watching BSL video (discussed in Section 2.5). EMT was applied to explore the visual response of deaf participants to BSL video stories in sequences which were selected to include a wide range of fine and gross movements and gestures. In this experiment, profoundly deaf adult participants, who used BSL as their first language, observed three short test BSL video sequences with the task of understanding the signed stories in each clip. Eye movement data were captured and compared for each subject and sequence. The EMT data were analysed with respect to the BSL content in the test video sequences to identify the impact of the cognitive task (understanding BSL) on visual behaviour.

6.4.2 Method

EMT apparatus was used to capture the gaze points of deaf subjects viewing three short BSL video sequences.

6.4.2.1 Subjects

The experiment was conducted with seventeen profoundly deaf-from-birth volunteers from the Aberdeen Deaf Social and Sports Club (ADSSC). These subjects were different from the volunteers who participated in the first EMT experiment (Section 6.3). For each subject, British Sign Language (BSL) was the first/preferred language. For this reason all communications were in BSL, aided by a local BSL Interpreter who was known to the participants.

Seven of the participants were excluded from the final results as it was not possible to obtain consistent accurate tracking of their eye movements during calibration, mainly due to head and body movements of the subjects and in one case due to interference from optical lenses worn by the subject. The problems associated with EMT calibration and poor success rates are discussed in Section 6.2. Of the ten subjects who completed the experiment, seven were male and three were female and their ages ranged from thirty to eighty-two years.

6.4.2.2 Apparatus

Eye movements were captured by a ViewPoint^{TM} eye tracker incorporating an infra-red light source and camera mounted on a clamp with a nose bridge and chinrest for comfortable and secure positioning of the subject's head (described in Section 6.1).

Test BSL video sequences were displayed to subjects on a seventeen-inch monitor (Monitor A) with true colour, 32 bit display connected to a Dell Pentium IV PC with PCI Video Capture Card installed. A second monitor (Monitor B) was connected to the PC (not visible to the subject) for the researcher to control and monitor the experiment (Figure 6.6).

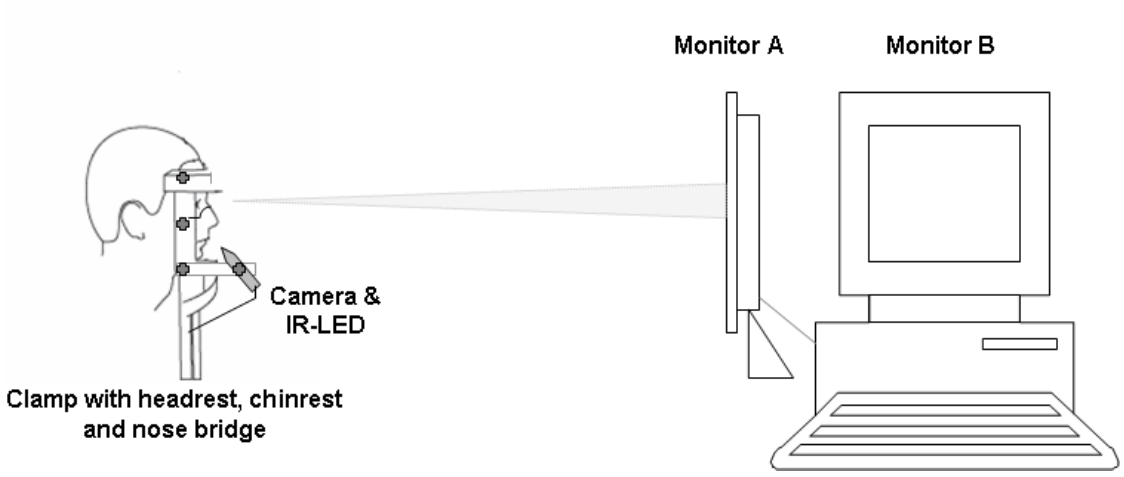


Figure 6.6 EMT Experimental Set-up

6.4.2.3 Materials

The BSL video material for the experiment was captured at 25 frames per second on a SonyVX200E Digital Video camera, under controlled artificial lighting in the university video recording studio, using two profoundly deaf volunteers (one male and one female). The volunteers were from the same geographical area, the North East of Scotland, and used the same version of BSL as the subjects participating in the experiment (the importance of this was discussed in Section 3.1). The signers in the video sequences related short stories from their own experience using their own natural style and expression of signing. Three video sequences were selected to ensure that the test material contained a wide range of sign language movements, expressions, and gestures (including finger spelling) described below.

Test BSL Video Sequence 1 (22.08 seconds) displayed a close (mid-shot) view of the signer (from the waist upwards). The female signer used facial expression, lip movement, and gestures but limited body movement around the scene which also included a camera and ventilation shaft. These background objects were included to test whether they would prove to be a distraction for the viewer. The story told in this clip is of the signer's experience of communication between deaf and hearing members of her family. An English translation of her story is:

"A long time ago, when I was young, I would ask my mother what everyone was saying. Now when my children speak with no voice, my mother asks me what they are saying. I remind her that she wouldn't tell me what was being said until they were finished and so she will just have to wait too. She realises this now".

In Test BSL Video Sequence 2 (27.20 seconds) the female signer (same as in Test BSL Video Sequence 1) was positioned at a greater distance from the camera and seen above knee height. She used facial expression, lip movement, wide gestures, and detailed finger-spelling but limited body movement around the scene which had no potentially distracting objects. The story told in this sequence was of the signer's experience as a child at school learning to use her voice. An English translation of her story is:

“When I was at school, a long time ago, when I was small, my speech was hopeless. They tried to teach me to speak but it just went over my head. The teacher said it was a bit of a problem. She said some people are good but you are not good, your speaking is not good. So I had to lie down on the floor and say ‘Ah’. She put a darning needle in my mouth to make the ‘Ah’ sound. My heart was throbbing.”

In Test BSL Video Sequence 3 (46.64 seconds) the male signer used facial expression, lip movement, finger-spelling, wide gestures, and movement around the scene to tell the story of his experience on holiday. An English translation of his story is:

“We have been to America, three times, and also to Spain. We met deaf people in America. The sign language was different but we could catch certain things by gesturing and so on. Things like ‘walking’, ‘hot’, ‘drinking’, and ‘good’ we could communicate, and also by writing things down. When I was a boy, I played football and so I could make conversation about that. The language was different, it was interesting.”

6.4.2.4 Procedure

Experiments were conducted under controlled conditions in a room with 100% artificial, overhead lighting. The subject was positioned at a comfortable viewing distance (four to six times the screen height) from Monitor A. Each subject was given instruction in BSL through a qualified Interpreter. All communication with the subject was in BSL, the subject’s first language. No printed instructions or feedback forms in English language were used.

The eye tracker camera was set up so that the video image of the subject’s pupil (dominant eye where appropriate) was in the centre of the control display window in Monitor B. The tracking system was adjusted in set-up mode (temporal resolution set to 30Hz, internal processing set to 340×240) so that the threshold area of the dark pupil of the eye and the white corneal reflection was located in the search area. The scan density was adjusted to obtain the minimum number of points which would correctly locate the dark pupil for maximum possible accuracy. Following the set-up stage, the equipment was calibrated for the individual subject to obtain coefficients for internal mathematical mapping calculations. Calibration was performed at temporal resolution

set to 30Hz, internal processing set to 640×480 to obtain the highest possible degree of accuracy. The subject was instructed to look at each of sixteen calibration points displayed on Monitor A, until they disappeared from the screen, avoiding anticipation of the next point. The researcher controlled and monitored the calibration routine on Monitor B, checking the success of calibration and re-presenting the stimuli as required. Once calibrated, the video stimuli (three video sequences separated by further calibration markers) were displayed full-screen to the subject on Monitor A. Eye movements were processed at temporal resolution set to 60Hz, internal processing set to 640×240 and monitored by the researcher on Monitor B. The (x, y) coordinates of the captured EMT data were saved to a unique data file for each subject.

The total time for the experiment with an individual participant was approximately twenty minutes. At the end of the experiment subjects were asked if there was anything in the sign language video that could not be understood, was not clear or needed to be repeated. The rationale for an open-ended question unrelated to the video content was that the experiment was designed to test ease of relaxed, natural sign language communication to the subject rather than test comprehension, which might have influenced the way the video material was viewed.

6.4.3 Results

All subjects reported ease of sign language communication with no requests for clarification or repetition. Conversations with the subjects after the experiment, through the BSL Interpreter, demonstrated understanding of and interest in the content of the video material used.

The EMT results were analysed for each participant by playing back the video clip and plotting the recorded (x, y) eye position coordinates on each video frame. The gaze points were examined frame-by-frame with respect to designated areas of the video image. The selected areas were: upper face, lower face, hands, fingers, upper body, lower body, background and object (a camera on a tripod in the first test video sequence). These were chosen so that the most important regions of the scene for BSL communication could be identified. The distinction between upper and lower face was made to determine if the region around the eyes (upper face) or around the mouth (lower face) was more significant for understanding BSL communication. The

distinction between hands and fingers was made to test whether wide movements of the hands and detailed movement of the fingers (for example, during finger spelling) were followed by the viewer. The upper body area was defined as the area below the chin and above the waist of the signer and the lower body was defined as the area below the waist. A fixation was recorded as a gaze of duration of 0.02 seconds or more (Palmer, 2002). In cases where the regions overlapped (for example, when the hands moved over the face region) the sequence of eye movements before and after this occurrence was observed to estimate which region was being followed by the eye.

The EMT data for each participant were analysed by determining the duration of fixations on designated regions of the image and also in relation to the BSL content in the test video sequences.

6.4.3.1 Fixation on Designated Regions of the BSL Video Image

The total fixation time on each of the designated video image regions was recorded to determine which region was most important to the viewer. In this approach, each subject's fixation time was expressed as a percentage of the total viewing time for each video sequence to allow comparison between viewers and to compare the results for each test.

The total fixation time (seconds) on the separate, previously described, designated regions of the video image was recorded for each of the ten subjects. Total fixation times for each subject varied depending on the number of saccades during viewing. The total and percentage fixation times for each subject, for each of the test video sequences, are given for each region of the video image in Tables 6.1, 6.2 and 6.3. These tables also show the average duration of gaze on each of the designated image regions. The average percentage fixation times are plotted in Figure 6.7 to allow comparison of the results obtained for the three BSL video sequences used in the experiment.

	Upper Face		Lower face		Hands		Fingers		Upper Body		Lower Body		Background		Object	
Subject	sec	%	sec	%	sec	%	sec	%	sec	%	sec	%	sec	%	sec	%
1	**21.44**	**97.99**	**0.28**	**1.28**	**0.00**	**0.00**	**0.00**	**0.00**	**0.04**	**0.18**	**0.12**	**0.55**	**0.00**	**0.00**	**0.00**	**0.00**
2	7.48	33.98	5.92	26.90	1.61	7.31	0.00	0.00	6.68	30.35	0.00	0.00	0.00	0.00	0.32	1.45
3	4.92	22.99	4.60	21.50	0.80	3.74	0.00	0.00	11.08	51.78	0.00	0.00	0.00	0.00	0.00	0.00
4	**21.56**	**98.54**	**0.00**	**0.00**	**0.00**	**0.00**	**0.00**	**0.00**	**0.32**	**1.46**	**0.00**	**0.00**	**0.00**	**0.00**	**0.00**	**0.00**
5	**21.52**	**98.35**	**0.36**	**1.65**	**0.00**	**0.00**	**0.00**	**0.00**	**0.00**	**0.00**	**0.00**	**0.00**	**0.00**	**0.00**	**0.00**	**0.00**
6	16.16	78.91	4.32	21.09	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
7	**21.08**	**96.17**	**0.84**	**3.83**	**0.00**	**0.00**	**0.00**	**0.00**	**0.00**	**0.00**	**0.00**	**0.00**	**0.00**	**0.00**	**0.00**	**0.00**
8	17.52	87.08	2.08	10.34	0.08	0.40	0.00	0.00	0.08	0.40	0.00	0.00	0.00	0.00	0.36	1.79
9	**18.78**	**90.03**	**1.64**	**7.86**	**0.44**	**2.11**	**0.00**	**0.00**	**0.00**	**0.00**	**0.00**	**0.00**	**0.00**	**0.00**	**0.00**	**0.00**
10	3.12	18.16	11.42	66.47	2.04	11.87	0.00	0.00	0.60	3.49	0.00	0.00	0.00	0.00	0.00	0.00
Average	15.36	72.22	3.15	16.09	0.50	2.54	0.00	0.00	1.88	8.77	0.01	0.05	0.00	0.00	0.07	0.32
St.Dev.	6.96	31.63	3.38	19.10	0.72	3.85	0.00	0.00	3.64	16.88	0.04	0.16	0.00	0.00	0.14	0.65
Average	**20.88**	**96.22**	**0.62**	**2.92**	**0.09**	**0.42**	**0.00**	**0.00**	**0.07**	**0.33**	**0.02**	**0.11**	**0.00**	**0.00**	**0.00**	**0.00**
St.Dev.	**1.06**	**3.21**	**0.58**	**2.76**	**0.18**	**0.84**	**0.00**	**0.00**	**0.12**	**0.57**	**0.05**	**0.22**	**0.00**	**0.00**	**0.00**	**0.00**

Table 6.1 Fixation on Different Regions of Test BSL Video Sequence 1

	Upper Face		Lower face		Hands		Fingers		Upper Body		Lower Body		Background	
Subject	sec	%	sec	%	sec	%	sec	%	sec	%	sec	%	sec	%
1	6.44	24.66	15.88	60.80	0.84	3.22	0.24	0.92	2.60	9.95	0.12	0.46	0.00	0.00
2	5.88	24.30	3.52	14.55	2.36	9.75	0.32	1.32	12.12	50.08	0.00	0.00	0.00	0.00
3	7.24	27.93	10.20	39.35	2.16	8.33	0.00	0.00	6.32	24.38	0.00	0.00	0.00	0.00
4	8.00	30.40	14.16	53.80	1.24	4.71	0.00	0.00	2.92	11.09	0.00	0.00	0.00	0.00
5	23.16	85.91	3.52	13.06	0.00	0.00	0.00	0.00	0.28	1.04	0.00	0.00	0.00	0.00
6	22.00	87.58	2.88	11.46	0.00	0.00	0.00	0.00	0.24	0.96	0.00	0.00	0.00	0.00
7	26.20	97.76	0.28	1.04	0.00	0.00	0.00	0.00	0.32	1.19	0.00	0.00	0.00	0.00
8	9.48	38.23	13.00	52.42	0.08	0.32	0.92	3.71	1.16	4.68	0.16	0.65	0.00	0.00
9	3.00	11.45	14.56	55.57	5.28	20.15	0.00	0.00	3.36	12.82	0.00	0.00	0.00	0.00
10	9.72	37.44	13.72	52.85	0.00	0.00	2.20	8.47	0.32	1.23	0.00	0.00	0.00	0.00
Average	12.11	46.56	9.17	35.49	1.20	4.65	0.37	1.44	2.96	11.74	0.03	0.11	0.00	0.00
St.Dev.	7.91	29.70	5.64	21.65	1.61	6.22	0.67	2.60	3.57	14.62	0.06	0.22	0.00	0.00

Table 6.2 Fixation on Different Regions of Test BSL Video Sequence 2

	Upper Face		Lower face		Hands		Fingers		Upper Body		Lower Body		Background	
Subject	sec	%	sec	%	sec	%	sec	%	sec	%	sec	%	sec	%
1	39.04	86.83	2.84	6.32	0.76	1.69	0.00	0.00	1.80	4.00	0.52	1.16	0.00	0.00
2	3.24	8.09	11.68	29.18	4.19	10.47	0.00	0.00	20.92	52.26	0.00	0.00	0.00	0.00
3	27.72	60.84	2.96	6.50	0.00	0.00	0.00	0.00	13.04	28.62	1.84	4.04	0.00	0.00
4	4.16	9.04	2.68	5.82	0.00	0.00	0.00	0.00	39.20	85.14	0.00	0.00	0.00	0.00
5	29.36	64.61	14.72	32.39	0.00	0.00	0.00	0.00	1.36	2.99	0.00	0.00	0.00	0.00
6	17.24	40.87	21.26	50.40	0.00	0.00	0.00	0.00	3.68	8.72	0.00	0.00	0.00	0.00
8	20.72	52.24	4.76	12.00	3.04	7.67	0.00	0.00	11.14	28.09	0.00	0.00	0.00	0.00
9	21.72	48.40	10.32	22.99	0.80	1.78	0.00	0.00	12.04	26.83	0.00	0.00	0.00	0.00
10	2.24	4.84	0.96	2.07	0.00	0.00	0.00	0.00	43.12	93.09	0.00	0.00	0.00	0.00
Average	18.38	41.75	8.02	18.63	0.98	2.40	0.00	0.00	16.26	36.64	0.26	0.58	0.00	0.00
St. Dev.	12.22	27.16	6.50	15.31	1.47	3.69	0.00	0.00	14.58	31.61	0.58	1.28	0.00	0.00

Table 6.3 Fixation on Different Regions of Test BSL Video Sequence 3

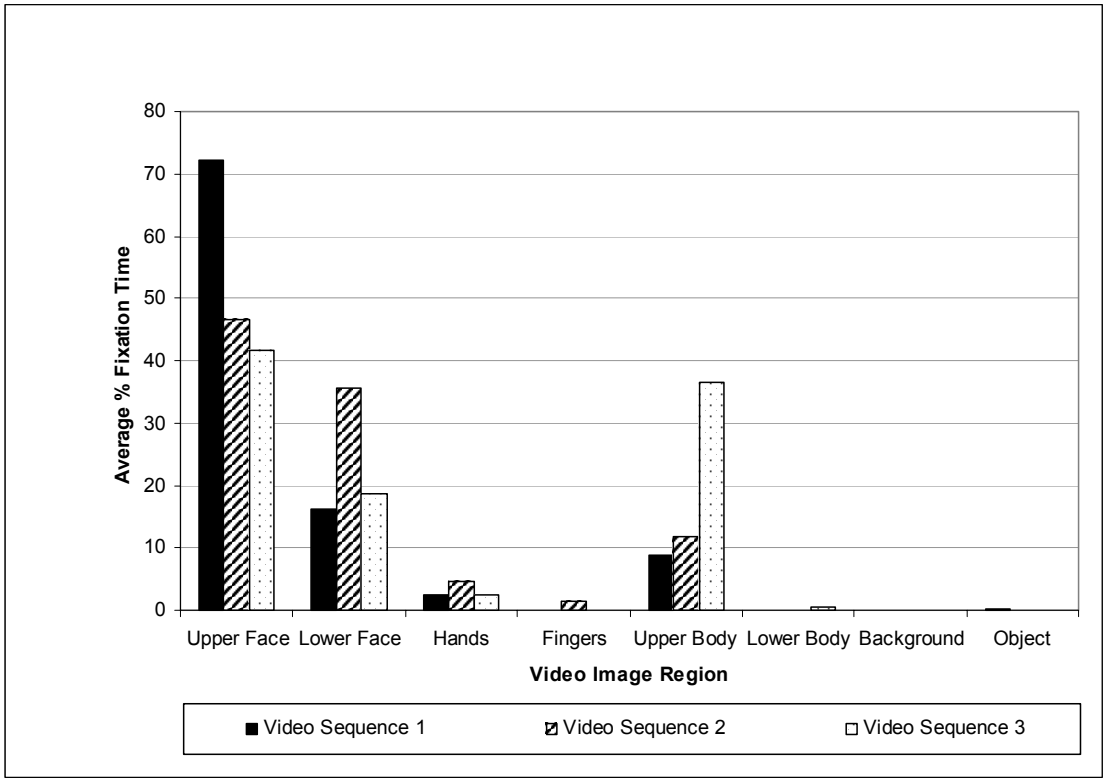


Figure 6.7 Average Percentage Fixation Times on Designated Regions of Importance for Test BSL Video Sequences 1, 2 and 3

The results for Test BSL Video Sequence 1 (Table 6.1) demonstrate that, on average, most of the time was spent looking at the face (88.31%) and, in particular, the upper face region (72.22%) of the video image. Subjects 1, 4, 5, 7, and 9 (shown in bold typeface in the shaded rows of Table 6.1) displayed a very similar pattern of viewing times and looked almost exclusively at the upper face during this video clip (96.22%). This is evident in the gaze coordinates for Subject 1, plotted on the video sequence in CD-ROM Appendix I. Subjects 6 and 8 exhibited behaviour similar to this group in terms of the time spent looking at the face, although their gaze fell on the lower face more than the rest of the group (21.09% and 10.34% respectively). The subjects spent an average of 0.5 seconds (2.54% of the total viewing time) looking at the hands and an average of 1.88 seconds (8.77% of the total viewing time) looking at the upper body region. The average fixation time on the lower body of the signer and the background object (camera) was less than the threshold time for a fixation.

The results for Test BSL Video Sequence 2 (Table 6.2) show that most of the fixation time (82.05%) was on the face region. Subjects 5, 6, and 7 exhibited similar behaviour to that for Test BSL Video Sequence 1 (an average of 90.42% of fixation time on the upper face). Subjects 1, 4, 8, 9, and 10 spent more time (an average of 55.09% fixation time) looking at the lower face region. Subjects 2 and 3 exhibited a similar viewing pattern to that shown for Test BSL Video Sequence 1; that is, fixating more on the upper body region.

Results for Test BSL Video Sequence 3 are shown in Table 6.3 (data for Subject 7 were excluded as he was the signer in the video sequence). The average time spent looking at the face in this test was 60.38% of the fixation time. More of the fixation time, 36.64% on average, was spent looking at the upper body region, which includes the area just below the face of the signer. Three of the subjects (Subjects 2, 4, and 10) spent most of their fixation time on the upper body.

The average fixation times on the designated areas, for each of the test BSL video sequences, (Figure 6.7) show patterns of viewing behaviour. The difference in the pattern of results obtained for the three BSL video sequences was explored further to determine if the data could have come from the same population (null hypothesis) or if the difference between at least two of the data sets was statistically significant.

6.4.3.2 Statistical Comparison of Viewing Behaviour

A non-parametric, Friedman Test (analysis of variance by ranks) was conducted to determine whether there was a statistically significant difference in the percentage fixation times on the specified regions of each of the three test BSL video sequences by the subjects in the sample at the 5% significance level. A non-parametric test is applied to ordinal or interval data, is distribution free and tests whether population locations differ (Hollander and Wolfe, 1999). This method was selected as it makes no assumptions about the frequency distributions of the variables being assessed. The EMT data (percentage fixation times) are interval data and not normally distributed. The null hypothesis for the test was that the data for all three test BSL video sequences could have come from the same population and were therefore not significantly different.

The Friedman Test ranks the results (percentage fixation times) for the subjects for each video and uses chi-square distributions to determine whether at least two of the data sets differ. The Friedman Test was conducted using SPSS (Statistical Package for the Social Sciences) and the output is given in Figure 6.8.

The test significance result gives a 0.097 probability that there was no significant difference in the results obtained for the three test BSL video sequences. This is greater than the level of significance (0.05 probability). The Friedman Test indicated that, for this sample, there was no statistically significant difference in the viewing behaviour of subjects for the three different video sequences used in the experiments.

Ranks

	Mean Rank
Video 1	2.56
Video 2	1.89
Video 3	1.56

Test Statistics[a]

N	9
Chi-Square	4.667
df	2
Asymp. Sig.	.097

a. Friedman Test

Figure 6.8 Friedman Test Results on EMT Data for Test BSL Video Sequences 1 to 3

6.4.3.3 Fixation in Relation to Video Content

Further examination of the raw EMT data was conducted to investigate the motivating factors for eye movements during the experiment. The sequence of fixations for each subject was examined with respect to the sign language content of each test BSL video sequence. A timeline was produced for each subject, which recorded the location of fixations during each of the video sequences with respect to video content (CD-ROM Appendix II). The data were examined on a frame-by-frame basis and the gaze point noted with respect to the BSL action in the video. Figure 6.9 is an extract from the timeline for Test Video Sequence 2. It shows the gaze locations of each of the ten

subjects for the first five seconds of the video sequence. BSL gestures are noted along the top row of the timeline and these are colour-coded to match the colours used to represent the designated regions of importance viewed in the video image in the figure. Sample frames from Test BSL Video Sequence 2 have been included in Figure 6.9 to illustrate the video image content at that point in the sequence.

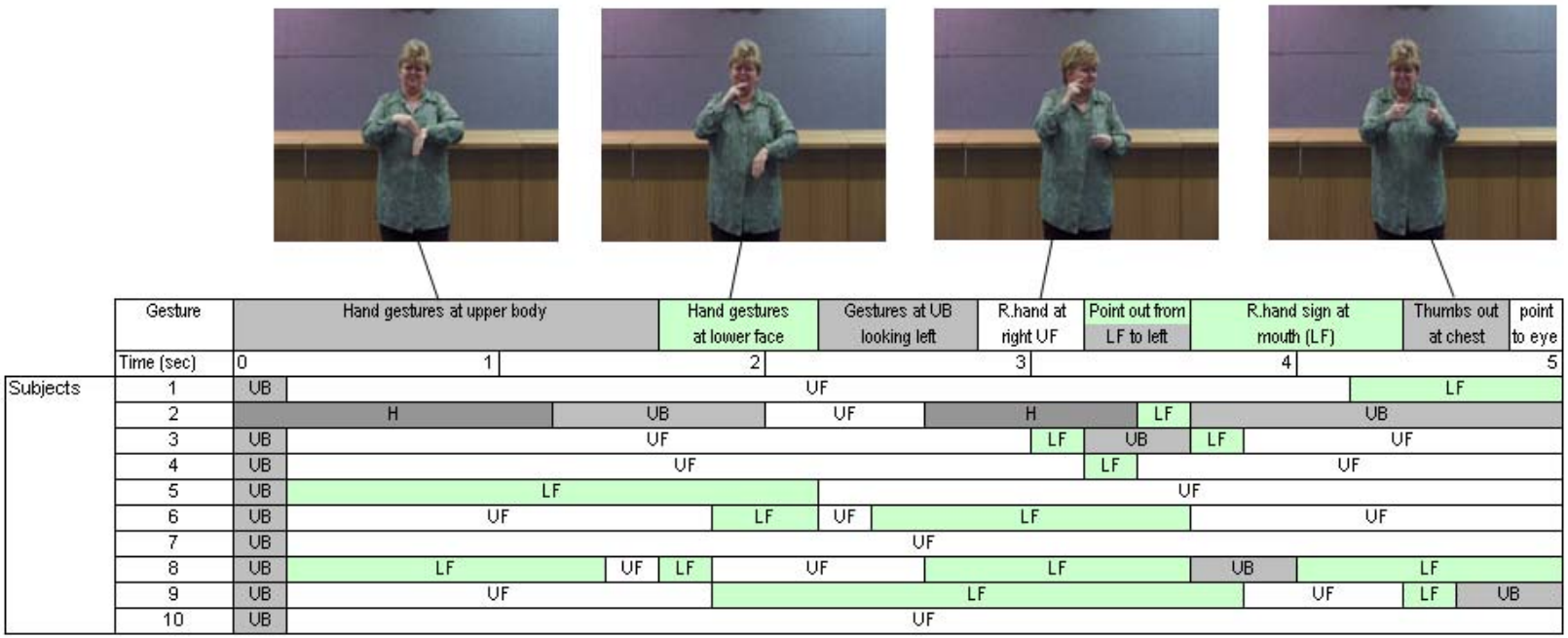


Figure 6.9 Extract from the Timeline Produced to Record the Location of Fixations of each Subject (1-10) with Respect to the Content of BSL Video Sequence 2

In Test BSL Video Sequence 1, the short visual excursions to the hands exhibited by Subjects 2, 3, 8, 9, and 10 were found to be associated with movement of the hands near (to one side of) the face region in the sign language video, possibly because the hands were close enough to “draw” the eyes away from the face but still allow the face to be seen in high (foveal) resolution. Two of the subjects (Subjects 2 and 3) spent a greater percentage of their total fixation time looking at the upper body region (30.35% and 51.78% respectively) which included the area just below the face. Examination of the timeline suggested that the gaze of these subjects was closer to the location of the hands than the other participants.

Motivating factors taking gaze away from the face in Test BSL Video Sequence 2 were investigated by examining the timelines for each subject. Gaze away from the face (mostly to the upper body region) occurred during pauses in the BSL and when gestures and movements were located in the lower body region of the signer. None of the subjects followed the hands or fingers during the periods of finger spelling in the video sequence. Gaze was found to be in the (upper or lower) face region during finger spelling in all cases.

Examination of the timelines for Test BSL Video Sequence 3 indicated that factors influencing gaze in the upper body region were large gestures (in the lower body region of the signer) and movement of the signer around the scene, particularly towards the end of the sequence.

The results determined that the face is the centre of attention for a deaf person observing sign language. Gaze was mostly in the upper face region for Test BSL Video Sequence 1 (in which there was a closer view of the signer) and more time was spent on the lower face in Test BSL Video Sequence 2 (in which the signer is further away from the camera, makes wider gestures and uses more finger spelling). Hand gestures close to the face, expansive gestures in the lower body region of the signer and movement of the signer around the video scene were found to act as motivating factors for taking the subject’s gaze away from the face region.

6.4.4 Discussion

The results of this EMT experiment support the conclusion that the most visually important region of the video image for BSL communication is the face of the signer in the video sequence. This was particularly evident in the results obtained for Test BSL Video Sequence 1, where the signer was closer to the camera than in the other test video sequences. Fixations were mainly on the upper-face region, with no visual excursions to the distracter objects in the background. Gaze was located mainly on the lower-face region for Test BSL Video Sequence 2 where the signer was further from the camera and the face region was therefore smaller. The wider, more rapid gestures and movements of the signer in Test BSL Video Sequence 3 seemed to cause gaze to fall more on the upper body region of the signer for some viewers than in the other test sequences. However, in each case the face was clearly the main region of importance and the Friedman Test determined that there was no statistically significant difference in the patterns of viewing behaviour across the three video sequences tested.

Analysis of the EMT data with respect to the BSL content in the test video sequences enabled the identification of conditions which caused eye movements to be directed away from the face.

From the EMT results, foveal resolution of the face region was found to be important for comprehension of sign language. Assuming that the hands of the signer play a significant part in sign language communication, it was postulated that they are observed in peripheral vision when they are not close enough to the face to be captured by the fovea of the eye. Peripheral vision was found to be adequate for the gross and rapid sign language movements of the hands and body that occurred away from the face region of the signer in the experiment. The deaf viewer fixates mostly on the facial region of the signer to pick up small detailed movements, associated with facial expression and lip shapes, which are known to convey important sign language information to the receiver. Small movements of the hands in front of or near to the face can be observed in the foveated region of view but more detailed movement near to the face was found to draw the eyes of some subjects away from the face for a short time. During this time the face was still close enough to be seen in high visual acuity. A deaf person uses peripheral vision to process information from larger, rapid

movements of the signer. Fixation on the upper body region (including the area below the face) by some subjects may have occurred to permit a range of smaller movements to be processed at the edge of the foveal area while still keeping the lower part of the face in high-resolution foveal vision.

6.5 Summary

The aim of the EMT experiments was to explore how profoundly deaf people view BSL video content for application in the design of systems for BSL communication. The impact of the visual task and the nature of the sign language material on gaze patterns were discussed in Chapters 2 and 3. The work of Siple (1978) was important for understanding the relationship between the Human Visual System and the development and production of sign language. The image content viewed in 'clear' foveal vision and the information that can be gathered from peripheral vision can be used to guide the development of systems (sign language or video communication systems) that work optimally within the limitations of human vision.

The EMT experiments presented in this chapter were designed to address methodological issues raised by previous researchers, successfully test the responses of deaf viewers to a range of sign language movements and gestures in 'real' BSL communication tasks and to investigate viewing patterns that might be exploited in the design of optimised video communication systems. The results provided new, world-first evidence that the most visually important region of the BSL video image is the face of the signer (Muir and Richardson, 2002) later confirmed in an independent study by Agrafiotis, Canagarajah, Bull and Dye (2003). Subjects were found not to track the movements of the hands or detailed movements of the fingers during periods of finger-spelling, suggesting that manual sign information was observed in peripheral (lower resolution) vision. Short visual excursions to the hands of the signer were evident only when they were close enough to the face for it to remain in foveal (high resolution) vision.

Further analysis of the image content of the BSL stimulus was conducted (Chapter 7) to explore what deaf viewers were foveating on and to determine how this content is coded in a standard video CODEC.

Chapter 7: Video Analysis

Eye Movement Tracking (EMT) experiments (Chapter 6) established a typical visual response to BSL video in deaf viewers. The most important region of the BSL video content was found to be the face of the signer. In addition to understanding the visual behaviour of BSL viewers, the content of BSL video images was analysed to determine the characteristics of the visually important regions and the regions observed in peripheral vision. It was also important to determine the relative demands of the different regions of the BSL video image content in terms of the number of bits required to encode them in a standard CODEC.

The objectives of the video analysis were to:

1. Investigate the optical flows in BSL video image sequences and identify the motion characteristics in the foveal and peripheral fields of vision.
2. Compare the number of bits required to encode different regions of the BSL video image content.

It was postulated that analysis of visually important video image content and how the BSL video image regions were encoded in a standard CODEC would identify opportunities for achieving higher compression ratios (and therefore, bit rate reductions) within acceptable visual limits for the task.

7.1 Optical Flow Analysis

Optical flow mapping techniques were applied to generate fields of image pixel trajectories. This facilitated analysis of the patterns of motion in the visually important regions of BSL video stimuli (in particular, the face region of the signer). The optical flows in the face region were compared to flows in the image regions which were not tracked by deaf viewers (that is, the hands and fingers of the signer). This section evaluates optical flow mapping techniques and describes the application of optical flow mapping to analyse the content of a sample of BSL video material.

7.1.1 Optical Flow Estimation

Motion perception is the process of inferring the speed and direction of objects that move in a visual scene (Section 2.4). The measurement of motion, or optical flow, in an image is fundamental in image processing. Motion information is essential in video compression (Section 4.1.3.2), computer vision and video analysis applications. Motion field analysis is important for real time video indexing and segmentation, event analysis and surveillance applications (Rapantzikos and Zervakis, 2005).

The HVS is capable of isolating objects in a complex scene using only motion information and so motion can be used as a cue for image segmentation. Motion-based segmentation can be applied as a pre-processing step in image processing, for example, face recognition and vehicle identification (CSIRO, 2001). Visual communications, such as sign language and lip reading, require motion processing (CSIRO, 2004), therefore investigation of image optical flows in the context of motion perception by deaf subjects was considered to be important in this thesis.

Digital video coding methods attempt to exploit the redundancy between consecutive or temporally close video images. Block-based video coding algorithms use motion estimation for compression purposes using block matching algorithms. The encoder transmits the location of the best matching blocks in the form of motion vectors (Section 4.1). The motion vector field is a crude approximation to optical flow. However, it may be heavily corrupted by noise caused by motion vectors not associated with real motion in the video scene due to the 'aperture problem' (Section 2.4) and illumination effects (Coimbra and Davies, 2005).

An optical flow map is a visual representation of object motion in a digital image sequence. It is typically a dense motion field with vectors at each pixel. Accurate techniques for estimating the optical flow field are important to many visual information applications (for example, pattern recognition, computer vision and other image processing applications).

Optical flow mapping techniques were applied to facilitate analysis of patterns of motion in the images in fixated and peripheral regions of the BSL video stimuli.

Horn and Schunck (1981) were first to develop an optical flow estimation technique based on computing spatio-temporal differences from image sequences. This technique is referred to as a 'global' method since it applies a global smoothness constraint to solve the 'aperture problem' (Section 2.4). An advantage of global methods (including the Horn-Schunck algorithm) is that they yield a high density of flow vectors; flow information missing in inner parts of homogeneous objects is filled in from the motion boundaries. A disadvantage is that this method is more sensitive to noise than alternative 'local' methods.

Local optical flow methods, such as the Lucas-Kanade (Lucas and Kanade, 1981) 'image registration' technique, attempt to calculate the motion between two image frames which are taken at times t and $t + \delta t$ at every pixel position. A pixel at location (x,y,z,t) with intensity $I(x,y,z,t)$ will have moved by δx, δy, δz and δt between the two frames. A disadvantage of local optical flow algorithms (including the Lucas-Kanade algorithm) is that they do not yield a very high density of flow vectors; the flow information fades out quickly across motion boundaries and the inner parts of large homogeneous areas show little motion detail. The advantage is comparative robustness in the presence of noise.

Barron, Fleet and Beauchemin (1994) conducted a quantitative evaluation of available methods. They reported results for nine optical flow techniques including differential, region-based matching, energy-based and phase-based methods. Differential techniques calculate velocity from spatio-temporal derivatives of image intensity or filtered versions of the image obtained using low-pass or band-pass filters. Region-based matching approaches define velocity as the shift that yields the best fit between image regions at different times by maximising a similarity measure. Energy-based methods (or frequency-based methods) are based on the output energy of velocity-tuned filters. Phase-based methods define velocity in terms of the phase behaviour of band-pass filter outputs. Barron *et al* (1994) compared the accuracy, reliability and density of velocity measurements obtained for real and synthetic image sequences and demonstrated significant differences in performance using these techniques. The conclusion of their study was that the most reliable and consistent methods were the differential method of Lucas and Kanade (1981) and the phase-based method of Fleet and Jepson (1990).

Galvin, McCane, Novins, Mason and Mills (1998) conducted an evaluation of eight optical flow methods, including six of the techniques evaluated by Barron *et al* (1994). Galvin *et al* (1998) concluded that, for real sequences with ground-truth data (obtained using a modified ray tracer), the technique by Lucas and Kanade (1981) gave the best results.

Accurate and efficient optical flow estimation continues to present a number of challenges which have been addressed by different researchers for specific applications. Rapantzikos and Zervakis (2005) addressed the need for robust, incremental, dense optical flow estimation to improve the optical flow field in MPEG sequences over a range of different motion scenarios as a basis for efficient video analysis. Coimbra and Davies (2005) aimed to overcome the reported problems of excessive averaging and error propagation to improve motion flow estimation in the MPEG-2 compressed domain.

Optical flows have also been calculated using motion blur (Section 2.4) information in still images (Rekleitis, 1996). However, if the motion blur is too small, it is undistinguishable from texture, noise or out-of-focus content and so this approach was not suitable for application to images in moving sequences in this thesis where small motions associated with facial expression were important.

Optical flow mapping techniques were applied to analyse motion flows in two test BSL video sequences. The methods were selected based on the review of available methods (Barron *et al*, 1994; Galvin *et al*, 1998) and with respect to the necessary trade-off between sensitivity to noise and density of flow vectors.

7.1.2 Method

Two optical flow methods were implemented; a simple 'global' optical flow estimation using correlation and a phase-based approach to the estimation of the optical flow field using spatial filtering (rather than by the spatiotemporal filtering approach of Fleet and Jepson (1990)) developed by Gautama and van Hulle (2002). These methods were selected from the preferred approaches of Barron *et al* (1994) and Galvin *et al* (1998).

7.1.2.1 Materials

Gautama and van Hulle (2002) found that performance of optical flow methods will "vary substantially for different image sequences". This fact highlights the importance of selecting appropriate video material for optical flow measurements which would be representative of the type of BSL video material communicated during a natural video telephony conversation. The selected test sequences contain a wide range of movements and expressions in a limited 'sign space' (that is, little movement of the signer in the scene) and with the signer displayed to the upper body region (designated image regions are described in Section 6.4.3). The video material for the optical flow experiments was chosen from a set of video materials created for subsequent subjective quality assessment by deaf viewers (Chapters 9 and 10). The creation of the original (source) video sequences is described in Section 9.2.3. Two test BSL video sequences were used in the optical flow experiments, described in Table 7.1.

BSL Test Sequence Number	Sequence Name (Duration)	English translation of BSL content
1	Lisa Family (10.16 seconds)	"I have one brother and one sister. I am older than them. My brother is divorced. He has two children."
2	Lisa Television (9.01 seconds)	"When I watch TV, I look at the subtitles or the signer at the bottom right and I look back and to the TV picture."

Table 7.1 Test BSL Video Sequences for Optical Flow Experiments

7.1.2.2 Correlation Method

A simple optical flow estimation technique using correlation was applied to BSL Test Sequence 1. The source video material was converted to 24-bit colour bitmap images for processing. The correlation algorithm, implemented in Matlab (Appendix C) converts two bitmap image files into luminance arrays for comparison of the Sum of Absolute Errors (SAE) in a predefined optimum search region (that is, a 5×5 pixel area) for flow density and computational efficiency. The SAE is a widely used measure of residual energy in a video image due to its computational simplicity (Richardson, 2003). The advantage of the correlation method is that it is a simple technique which provides a high density of flow vectors. However, this is at the expense of sensitivity to noise.

This method was chosen since noise was not a significant problem in the source material and high density of flow vectors was desirable for video analysis, particularly for small motion flows.

7.1.2.3 Phase-Based Method

A phase-based approach to the estimation of the optical flow field, based on the approach of Fleet and Jepson (1990) and developed by Gautama and van Hulle (2002), was applied to Test BSL Sequence 2. Open source Matlab code was obtained for the Gautama and van Hulle (2002) phase-based optical flow algorithm (available from the Matlab Central File Exchange, http://www.mathworks.com/matlabcentral/fileexchange). It was developed for processing video images (CIF files) and the input parameters: video file name, number of frames, number of evaluation points on the x-axis (gx) and the temporal span (ts) (Appendix D). This method plots the estimated flow vectors on the first input frame and was selected as it enables the visualisation of large motion flows in the video images.

7.1.3 Results

The optical flow methods were applied to the test BSL video sequences and optical flow maps were produced for analysis of motion with respect to the BSL video image regions.

Figures 7.1 and 7.2 illustrate a sample of the results of the correlation technique applied to Test BSL video sequence 1. These examples reveal high density flows in the foreground object (the signer) and minimal noise in the background region. They demonstrate the general observations made for different test portions of the video sequence; firstly, where there is action in the arms/hands/fingers region and little movement in the face region (Figure 7.1) and secondly, where there is action in both of these regions (Figure 7.2). The correlation technique provided detailed information about motion flows with little interference due to the limited amount of noise in the source material. It was possible to observe the location of small optical flows without superimposing the flow vectors on the video frames.

The optical flow maps obtained by the phase-based method are represented by the examples given in Figure 7.3. The temporal phase gradients are plotted for the first 9 frames of test BSL Video Sequence 2 over a temporal span (ts) of 3 (Figure 7.3a) and temporal span (ts) of 5 (Figure 7.3b). A lower temporal span produces a higher density of vectors. The number of evaluation points in the horizontal dimension (gx) was set to 25. This was the value determined in the algorithm by Gautama and van Hulle (2002) for producing distinct flow vectors. Analysis of the results for the remainder of the frames in the test BSL video sequences was conducted using the same parameters as those for Figure 7.3(b). The phase-based method produced fewer vectors over a greater temporal span than the correlation method and so it was possible to observe the large motion vectors in the BSL video images.

Detailed analysis of the results for each video revealed that the optical flows in the face of the signer were small/fewer compared to the flows in the other, less visually important regions, of the image for sign language users. Large motion vectors were identified in less visually important regions of the image; that is, the hands and arms of the signer which are not tracked by deaf viewers. It was determined that the deaf viewer foveated on the face of the signer to receive the fine motions of facial expression (known to be essential for understanding the meaning of BSL gestures) and that gross sign language movements were observed in low resolution peripheral vision. The next section of this chapter explores how the motion flows in the visually important regions of the image and the regions viewed in peripheral vision are encoded in a standard CODEC.

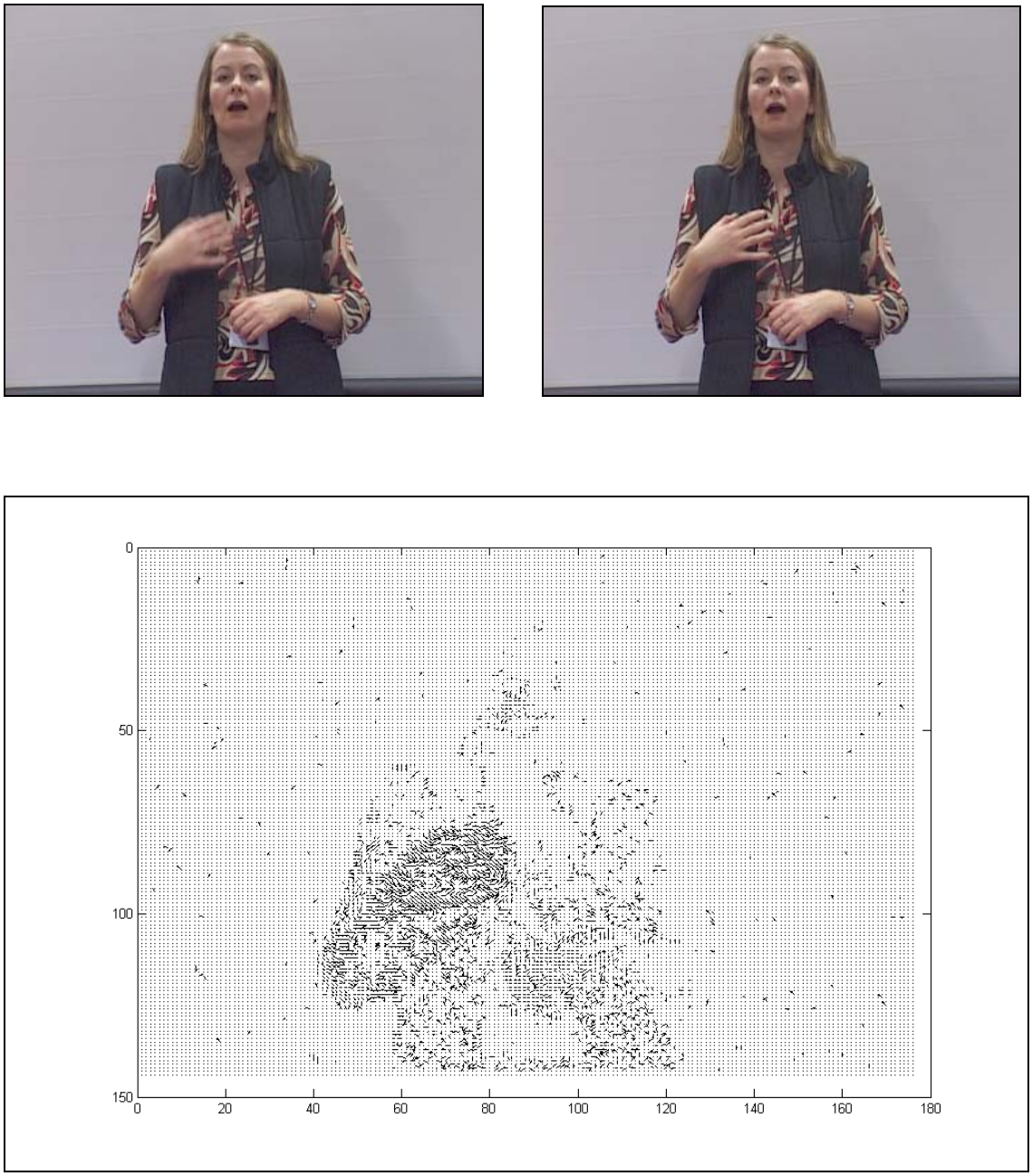


Figure 7.1 Test Video Sequence 1 (Lisa Family) Frames 22 and 23 and the Resulting Optical Flow Map Generated in Matlab by the Correlation Method.

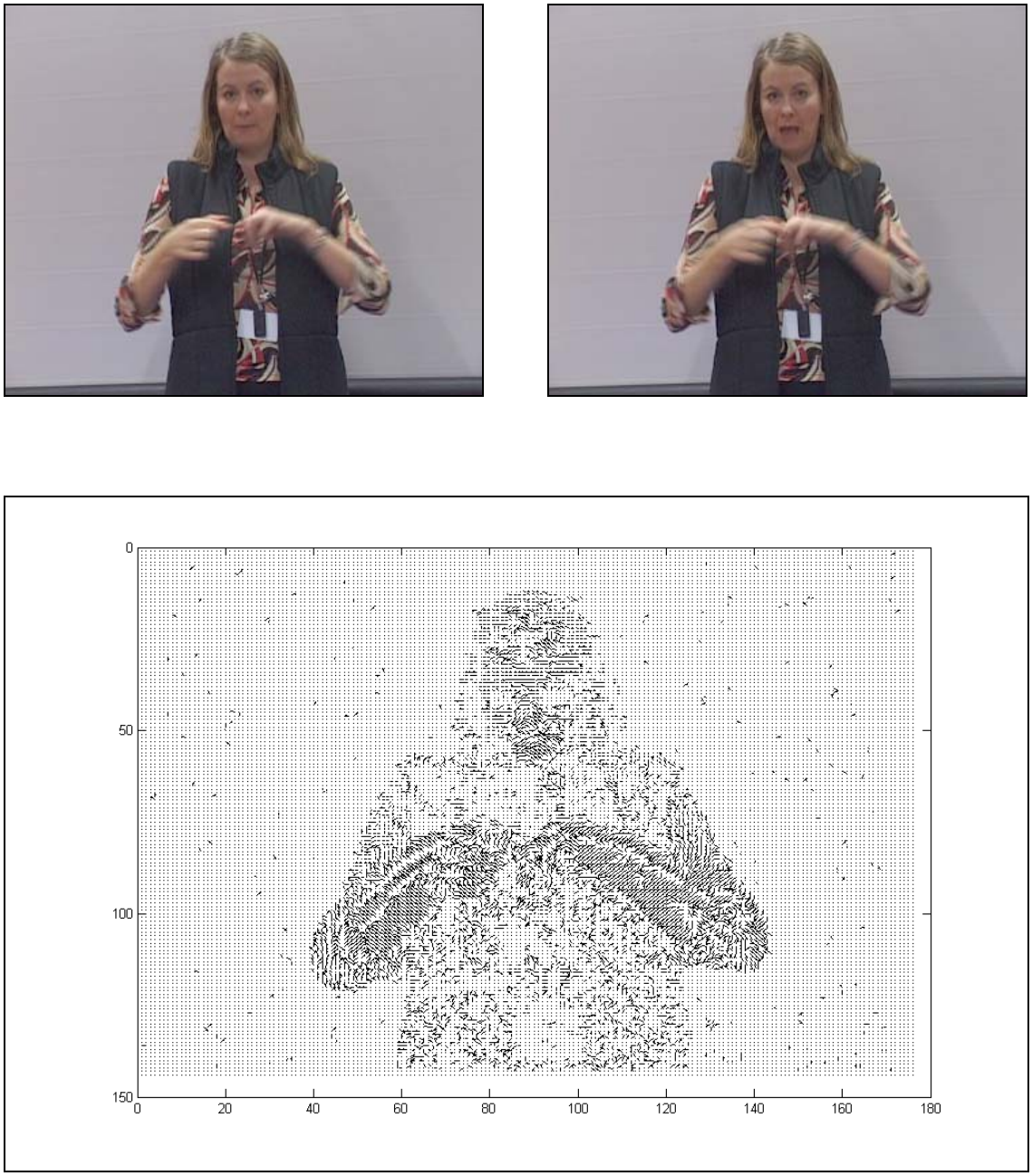


Figure 7.2 Test Video Sequence 1 (Lisa Family) Frames 201 and 202 and the Resulting Optical Flow Map Generated in Matlab by the Correlation Method.

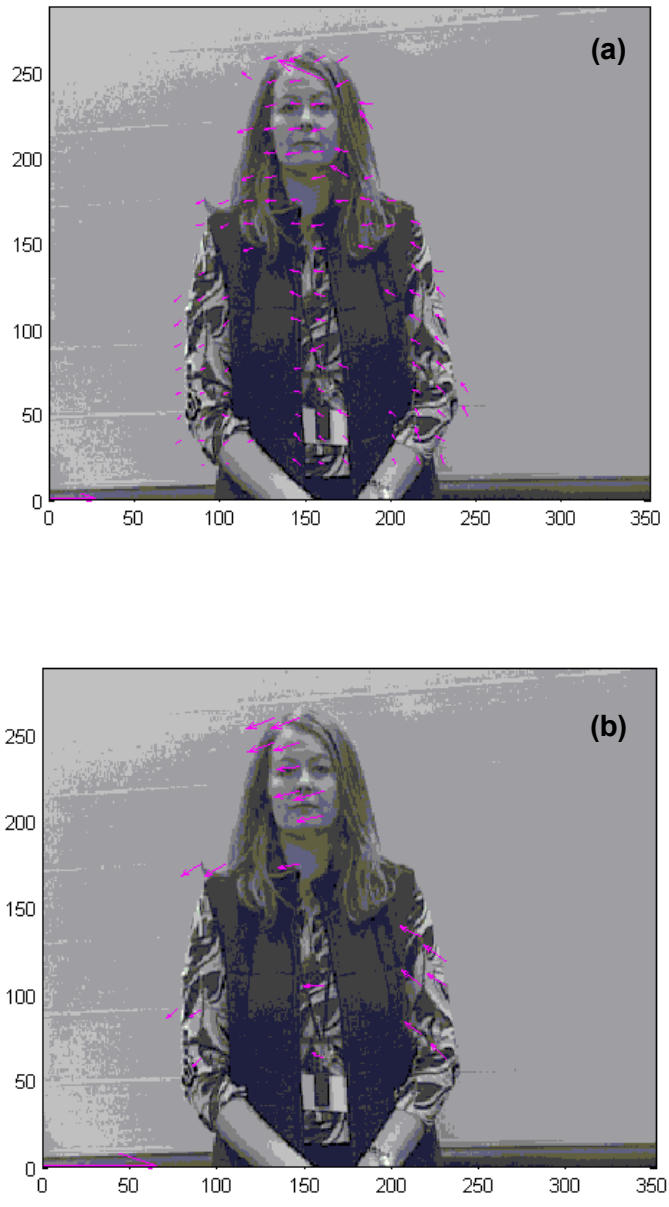


Figure 7.3 Optical Flows Generated in Matlab by the Gautama and van Hulle (2002) Phase-Based Method for Test Video Sequence 2 (Lisa Television) (a) 9 frames, ts = 3, gx = 25 (b) 9 frames, ts = 5, gx = 25.

7.2 Encoded Bit Count Analysis

Encoded bit count analysis was conducted to determine which regions of BSL video images were given highest priority (in terms of the number of bits used to encode the macroblocks in the image regions) in a standard coding scheme.

7.2.1 Method

Six BSL video sequences (Table 7.2) were encoded using a basic H.263 encoder (equivalent to MPEG-4 Simple Profile). The test BSL video sequences (CIF format) were selected for encoded bit count analysis from the set of video materials used in the EMT experiments (described in Section 6.4.2.3) and from those created from source material (described in Section 9.2.3) for subsequent subjective quality assessment by deaf viewers. The total number of encoded bits in each macroblock position of the sequences was calculated from data extracted from the encoder trace file using an algorithm implemented in Matlab.

BSL Test Sequence Number	Sequence Name (Duration)	English Translation of BSL content
1	KathSeq1 (22.08 seconds)	"A long time ago, when I was young, I would ask my mother what everyone was saying. Now when my children speak with no voice, my mother asks me what they are saying. I remind her that she wouldn't tell me what was being said until they were finished and so she will just have to wait too. She realises this now".
2	KathSeq2 (27.20 seconds)	"When I was at school, a long time ago, when I was small, my speech was hopeless. They tried to teach me to speak but it just went over my head. The teacher said it was a bit of a problem. She said some people are good but you are not good, your speaking is not good. So I had to lie down on the floor and say 'Ah'. She put a darning needle in my mouth to make the 'Ah' sound. My heart was throbbing."
3	Lisa Family (10.16 seconds)	"I have one brother and one sister. I am older than them. My brother is divorced. He has two children."
4	Lisa Introduction (10.18 seconds)	"Hello, my name is Lisa. My dog's name is Bran. He is a hearing dog for the deaf. He helps me."
5	Lisa School (8.03 seconds)	"I went to Aberdeen School for the Deaf. As I was growing up I used signs and learned oral communication."
6	Lisa Television (9.01 seconds)	"When I watch TV, I look at the subtitles or the signer at the bottom right and I look back and to the TV picture."

Table 7.2 Test BSL Video Sequences for Encoded Bit Count Analysis

7.2.2 Results

The total number of encoded bits at each macroblock position (for 18×22 macroblocks in the CIF sequences) was plotted for each test BSL video sequence (Tables 7.3a and b). A single (sample) frame is given in the tables to illustrate the approximate positions of the macroblocks relative to image content. The encoded bit maps illustrate the

regions which require the highest number of bits to encode them in red and the fewest in dark blue.

The array of total bit counts for each macroblock in the sequences, obtained from the encoder trace file data, was analysed with respect to different designated regions of the BSL video images. The six designated image regions for analysis were similar to the regions used to evaluate the EMT data (Section 6.4.3), that is the background, background object (Test BSL Sequence 1 only), head/face, arms/hands, upper body and lower body. The total bit counts for each region were calculated to account for the side-shifts in body motion and head movements in the sequences. An example of the total bit count map obtained is given in Figure 7.4 which shows the colour-coded total bit counts for each of the designated video image regions in BSL Test Sequence 1. The total bit count maps for BSL Test Sequences 1 to 6, including the calculations for each of the designated image regions, are given in CD-ROM Appendix III. In situations where the arms/hands overlapped the upper and lower body regions, the bit count was included in the total for the arms/hands. The total number of bits used to encode each of the designated video image regions for each test BSL video sequence is given in Table 7.4.

The macroblocks in the image regions corresponding to the arms and hands of the signer in the BSL video sequences were encoded with the highest number of bits (43 – 67% of the total bit count for the sequence). The total number of bits used to encode the head/face region (where activity was relatively low) represented 8-11% of the total bit count for the BSL video sequence.

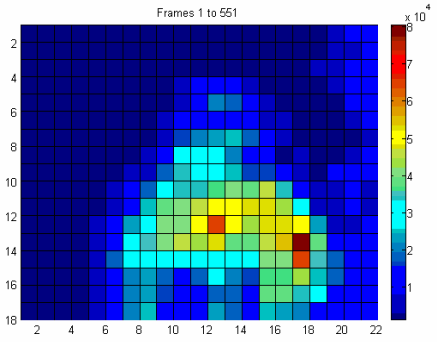
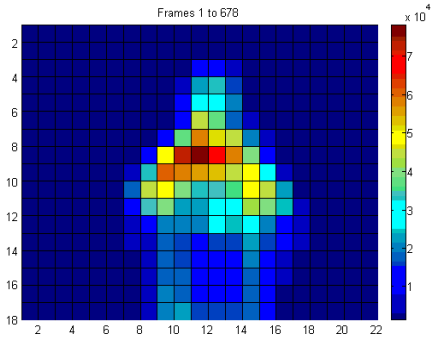
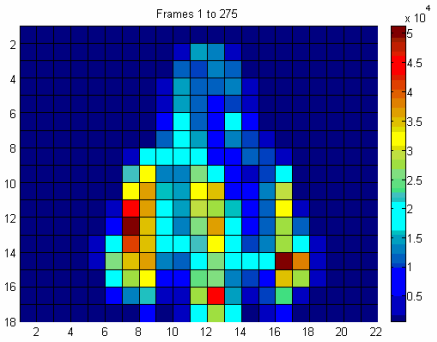
BSL Test Sequence Number	Macroblock bit count density map	Sample Video Frame
1		Frame 443 
2		Frame 536 
3		Frame 23 

Table 7.3a Encoded Bit Count Maps for each Macroblock of Test BSL Video Sequences 1 to 3 in CIF Format

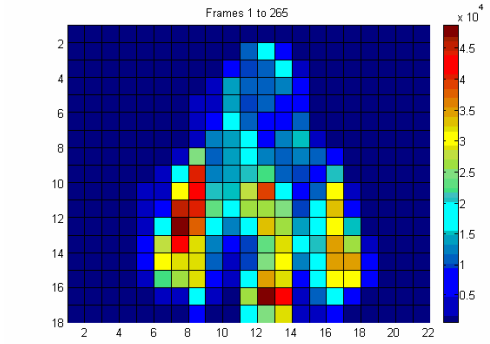
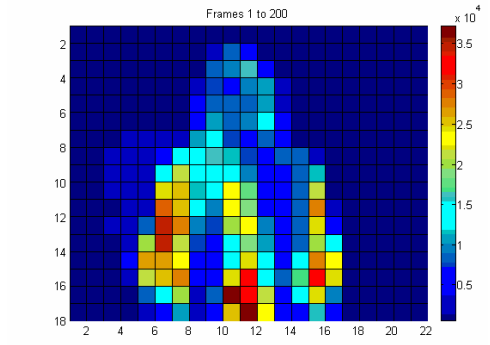
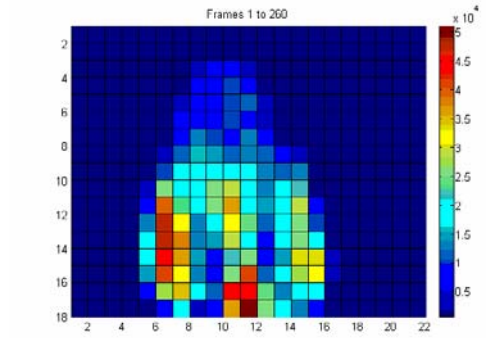
BSL Test Sequence Number	Macroblock bit count density map	Sample Video Frame
4		Frame 80 
5		Frame 41 
6		Frame 175 

Table 7.3b Encoded Bit Count Maps for each Macroblock of Test BSL Video Sequences 4 to 6 in CIF Format

MBs	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22
1	613	731	674	720	758	729	725	697	724	711	740	702	728	674	690	691	2127	2144	4311	11584	13128	13987
2	664	713	773	731	734	693	750	742	786	741	707	731	705	724	681	677	2358	2200	5150	7060	13242	15136
3	695	715	717	717	701	766	783	694	756	702	678	657	704	764	787	651	1374	4101	5026	9880	12996	14886
4	748	726	757	690	711	732	718	748	815	4185	9342	11317	8126	3733	1986	1565	1062	2072	3768	8347	12457	14006
5	755	753	724	781	781	712	698	866	2928	7784	13868	19769	18883	8971	4348	2494	692	2923	2462	8803	12637	13427
6	727	771	783	730	757	711	746	972	5578	9418	9933	13432	15077	10218	5774	3599	590	850	1515	5545	9744	12923
7	705	736	753	723	710	692	758	1579	7410	15784	20235	23219	24806	16940	8275	3903	815	629	619	1572	8013	12364
8	705	722	726	743	741	849	1329	3572	9687	29831	26860	29140	19474	10044	11220	2496	4594	1533	2625	7528	10998	12354
9	691	755	766	806	991	1276	3563	7975	17569	34105	28853	28801	25307	15749	16436	11048	4920	2999	9077	8698	10717	11747
10	696	735	717	926	1501	3482	8647	19140	29678	33710	33919	36203	39639	38898	44825	33364	14340	3604	6037	9957	7279	6964
11	721	721	727	1042	2705	5670	15467	25351	37563	39577	46179	50719	50770	47443	52593	44400	25861	6775	6277	8730	5963	6615
12	698	738	727	1242	3454	8295	23229	30996	39935	39923	49315	64352	50014	43070	52211	52840	48313	23346	7563	7823	10093	11720
13	717	699	725	1070	4320	14598	31951	34727	40482	44704	43508	48088	37835	39012	46851	53476	80940	38459	14241	11828	10080	11272
14	732	693	773	927	5239	18741	28299	29067	29778	29373	30260	31506	29758	30398	41270	41982	63534	35289	17214	14850	12237	11385
15	724	745	792	940	4818	13494	20207	23742	17660	14638	19422	29439	21259	15967	36027	36478	42693	24695	18125	14878	13426	11684
16	721	692	724	1076	3729	10071	19942	22044	15942	11162	12876	18201	18334	13016	38118	38667	33814	28345	15583	11002	14824	11607
17	669	725	777	1340	3543	10387	21182	24648	15072	11296	11708	22840	18670	12546	24492	23192	20317	12212	8601	8169	11223	9623
18	987	1211	1122	2136	1751	3417	11870	19681	16162	12801	15612	18950	13942	10686	21770	23364	16294	8302	8581	9351	5617	9810

Background	Object	Head/Face	Upper Body	Arms/Hands	Lower Body

Figure 7.4 Colour-Coded Total Bit Count Array for BSL Test Sequence 1

Sequence Number	Background		Head/Face		Arms/Hands		Upper Body		Lower Body		Object		TOTAL	
	Bit Total	%	Bit Total	%	Bit Total	%	Bit Total	%	Bit Total	%	Bit Total	%	Bit Total	%
1	145833	3.01	366413	7.56	2564807	52.91	691888	14.27	374557	7.73	703872	14.52	4849866	100
2	211472	6.90	310269	10.13	1485222	48.47	539528	17.61	517697	16.90	N/A	N/A	3064188	100
3	112590	4.26	285661	10.82	1771039	67.08	235745	8.93	234999	8.90	N/A	N/A	2640034	100
4	109649	3.93	262774	9.42	1238055	44.37	506263	18.14	673545	24.14	N/A	N/A	2790286	100
5	119257	5.47	219370	10.07	929496	42.65	460285	21.12	450928	20.69	N/A	N/A	2179336	100
6	118879	4.65	232694	9.09	1711836	66.90	326959	12.78	168280	6.58	N/A	N/A	2558648	100

Table 7.4 Total and Percentage Bit Count for each of the Designated Video Image Regions in the Test BSL Video Sequences

7.3 Discussion

Video analysis, using optical flow and encoded bit count estimation techniques, identified the motion characteristics and relative encoding requirements of designated regions of BSL video image content.

The results of EMT experiments (Chapter 6) determined that the most visually important region of the BSL video image was the face of the signer. Optical flow analysis of BSL video revealed that motion flows were greatest in the regions that included movement of the arms and hands, which are observed by deaf viewers in peripheral vision. Small motion flows in the face region (associated with small rapid changes in contrast in the images) were evident in the foveated field of view. This suggested that the design of a video communication system for deaf people, which gave priority to the face of the signer, would need to preserve some of the higher frequency components of the image in the face region.

Analysis of the number of bits required to encode different regions of the BSL video image content revealed that coding priority (in terms of bit count allocation) was given to the region containing the largest motion vectors, that is the arms/hands region which was allocated 43-67% of the total bit count. The total bit count for the head/face region was found to be 8-11% of the total for the sequence which was less than the percentage bit count for the stationary background object in Test BSL Sequence 1 (14%) and not much more than the percentage bit count for the background region in all sequences (3-7%).

This is the expected result for standard coding of high motion content. A standard 'lossy' video CODEC must achieve compression at the expense of image quality (Section 4.1.3). High frequency spatio-temporal components (representing areas of significant motion and/or residual detail) in a video image require more bits for encoding (as was demonstrated in Section 7.2.2 for the arms/hands region of BSL video). This is due to the relatively high number of DCT coefficients (specifically, variable length codes) and differential motion vector components which contribute significantly to the overall bit output of the encoder in high motion sequences (Sadka, 2002). There is significant motion in the arms/hands regions of the BSL video images, generating many motion vectors. In addition, the irregular object shapes (of the hands

and fingers) are difficult to compensate for accurately, so there is significant data remaining in the residual.

Standard systems produce a variable bit rate for a constant QP during encoding to maintain constant perceptual quality in the decoded sequence. A system operating within the constraint of limited bandwidth at a QP fixed to compensate for large movements in a high motion sequence will typically remove the finer spatio-temporal detail in the image. Thus, a CODEC designed to match limited bit rate will allocate most of the bit budget to motion vector and residual changes in areas of significant motion and this has the effect of giving coding priority to these areas of the video image.

Current video coding standards, which give priority to low spatial frequencies (changes in luminance over a large area associated with larger motion flows) in image content at a fixed QP for limited bit rates, do not match the visual requirements of BSL users. Visually important content regions for BSL users are not generally those regions with large magnitude optical flows. BSL users have been shown to give priority (via foveal vision) to important high frequency information in the face of the signer in BSL video. Therefore more bits are used to encode less important regions of the image for BSL communication in standard video CODECs at limited bit rates. This anomaly strengthens the case for design of systems based on actual viewing behaviour for a visual task; the approach in this thesis.

It was postulated that, by changing the priority given to coding different image content regions in BSL video to meet the visual requirements and sensitivities of deaf viewers (specifically, force the encoder to 'spend' more of the bit allocation on the face region), the bit allocation for some of the high frequency components in the encoded images could be reduced. This would result in a potential saving in transmission bandwidth and better perceived quality for the viewer. This hypothesis was tested by pre-processing BSL video material using methods developed to give priority to image quality in the face of the signer (Chapter 8) and obtaining feedback on the quality of the outputs from BSL users (Chapter 9).

Chapter 8: Content-Prioritised Video Coding

Eye movement tracking experiments (Chapter 6) established that BSL users exhibit a characteristic visual response to BSL video content. Deaf viewers fixate mostly on the facial region of the signer in the test BSL video sequences to pick up small detailed motion (Section 7.1) associated with facial expression and lip/mouth shapes which are known to be an essential part of the signed communication. Eye movements direct the fovea (Figure 2.1) of the eye (which is responsible for high resolution vision) to the face of the signer. Peripheral, low resolution vision was found to be adequate for perceiving the gross and rapid gestures that occurred away from the face region of the signer.

The aim of the experimental work in this chapter was to develop methods of content prioritisation of BSL video which exploited the visual behaviour of deaf people determined by EMT and the properties of foveated vision. The hypothesis was that, by giving coding priority to the most important content of the video image for BSL communication using a model of HVS foveation, best quality for the user task would be achieved for a higher degree of compression and reduced bandwidth requirements.

Novel spatial and temporal video image foveation methods are described which pre-process BSL video image content based on the foveal response of the HVS and the visual response of deaf people to BSL video communication.

8.1 Spatial Video Image Foveation Method

The spatial video image foveation algorithm developed for this thesis was based on the multi-resolution pyramid model described by Geisler and Perry (1998 and 1999) and Kleinfelder (1999), discussed in Section 2.3.2. The algorithm was developed in Matlab (Appendix E) to process BSL video sequences and permit variable spatial contrast threshold input.

In the spatial video image foveation algorithm, the BSL video image data are decomposed into a pyramid of 2-D arrays of coefficients representing different spatial frequency bands. The first level of the seven-layer pyramid contains the greatest number of coefficients and highest spatial frequency band. Each successive level contains one-quarter of the number of coefficients of the previous level. The input

image (Level 1) is low-pass filtered and then down-sampled by a factor of two in both directions to obtain a lower resolution image (Level 2) with one-quarter of the number of elements. Low-pass filtering and down-sampling is repeated to obtain a sequence of successively lower resolution images. Up to seven levels are computed in the foveation model (three levels are illustrated in Figure 8.1). The schematic diagram in Figure 8.1 shows the outer boundary of the foveation regions at each level in the low-pass pyramid represented by the inner solid squares in the top row. The inner boundaries of the foveated pyramid regions are indicated by the dashed squares in the top row. The inner and outer boundaries of the foveated pyramid regions are also illustrated in the bottom row of the diagram. The shaded regions indicate the image elements which are processed further and the overall result of the process is that the amount of image data which needs to be processed is significantly reduced.

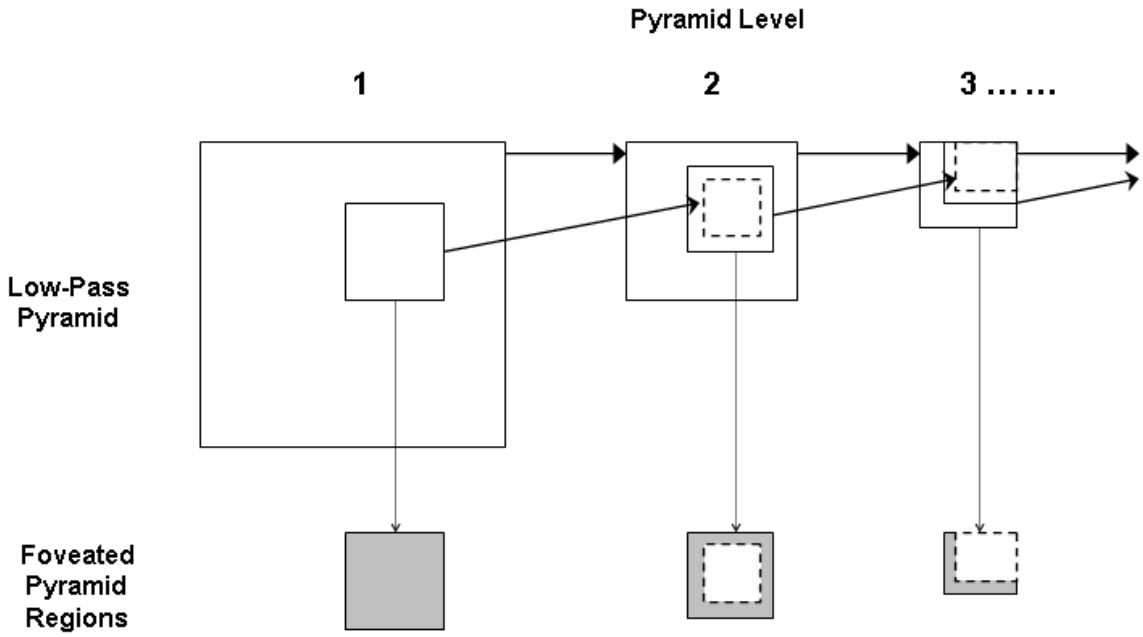


Figure 8.1 Computation Process for the Foveated Multi-Resolution Pyramid (Adapted from Geisler and Perry, 1998)

Blending and foveation-point interpolation produces a smooth transition between the levels of the pyramid and artefact-free foveation of the images. The pyramid levels are illustrated in Figure 8.2.

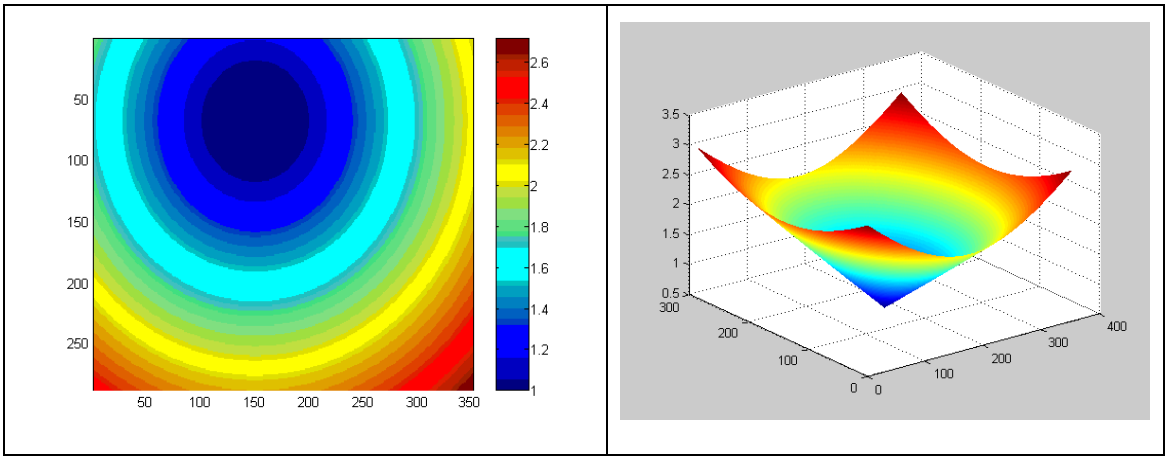


Figure 8.2 2-D and 3-D Plots of Video Image Foveation Maps

The dark blue region in the video image foveation maps represents the foveated ('clear') region corresponding to the face of the signer in the video image. This is Level 1 in the multi-resolution pyramid. The fractional values in the scale for the plots are constrained in the algorithm to conform to the 'Levels' of the multi-resolution pyramid. Up to seven levels are produced in the image foveation algorithm. The actual number of levels computed (three levels are produced in the example in Figure 8.2) depends on the size of the central foveated region which is determined by the viewing distance from the image display. The different foveation regions (or Levels) are determined using a contrast threshold function (Equation 8.1) based on human contrast sensitivity data measured as a function of spatial frequency and retinal eccentricity.

$$CT(f,e) = CT_0 \exp(\alpha f (e + e_2)/ e_2) \qquad \textit{Equation 8.1}$$

In Equation 8.1, CT_0 is the minimum contrast threshold, α is the spatial frequency decay constant, e_2 is the half-resolution eccentricity (in degrees) at which visual acuity is half as good as at the centre of the fovea and f is the maximum spatial frequency discernable at a given retinal eccentricity e (in degrees).

This formula fits with published contrast sensitivity data for small, briefly presented patches of grating (Robson & Graham, 1981). Matching the foveation to the fall-off in resolution of the HVS with eccentricity makes optimal use of foveation because it removes image information which can not be resolved. The parameters in the spatial video image foveation algorithm are given in Table 8.1.

Viewing distance	smaller value gives bigger 'clear' area (variable)
Alpha (α)	spatial frequency decay constant (set to 0.106)
Epsilon2 (e_2)	half resolution eccentricity (set to 2.3)
CT_0	minimum contrast threshold (set to 0.0133)

Table 8.1 Foveation Parameters and Psychophysically Measured Values for the Geisler and Perry (1998) Image Foveation Method

The spatial video image foveation algorithm is applied to image data in each frame of the video sequence. The main processing stages in the algorithm are described (with Matlab code[14]) as follows:

1. The x and y off-sets of each pixel (ex and ey) compared to the point of focus (fovx, fovy) in pixels are computed:

   [ex, ey] = meshgrid(-fovx+1:img_size(2)-fovx,-fovy+1:img_size(1)-fovy);

2. The data in Table 8.1 and the x and y off-sets of each pixel (ex and ey) are used to calculate a mesh (ec) the size of the original video image. The radial distance (in meters) between each point and the point of gaze (eradius) is computed based on the viewing distance (Table 8.1) and the display dot pitch[15].

   eradius = dotpitch . sqrt(ex.^2+ey.^2);*
   *ec = 180*atan(eradius ./ viewingdist)/pi;*

3. The maximum frequency (maxfreq), in cycles per degree, of the display device (for the viewing distance set in the algorithm) is computed:

---

[14] The following Matlab operators/functions are applied in the algorithm: **.*** performs array multiplication; for example, A.*B is the element-by-element product of the arrays A and B. **atan** computes the inverse tangent in radians; for example, Y = atan(X) returns the inverse tangent (arctangent) for each element of X. **meshgrid** is applied as follows, [X,Y] = meshgrid(x,y) transforms the domain specified by vectors x and y into arrays X and Y.

[15] Dot pitch is a specification for a computer display that describes the distance between phosphor dots (sub-pixels) or LCD cells of the same colour on the inside of a display screen. A smaller number generally means a sharper image (as there are more dots in a given area), and vice versa. A typical monitor has a diagonal dot pitch of 0.28 mm; a good quality monitor has a diagonal dot pitch of 0.26 mm.

```
maxfreq = pi ./ ((atan((eradius+dotpitch)./viewingdist) - ...
        atan((eradius-dotpitch)./viewingdist)).*180);
```

4. The contrast threshold (Equation 8.1) is set to a maximum value of 1 to obtain a matrix form of f (spatial frequency in cycles per degree at each pixel) in the formula computed as eyefreq in the algorithm.

```
eyefreq = ((epsilon2 ./(alpha*(ec+epsilon2))).*log(1/CT0));
```

5. The pyramid level (pyrlevel) is the fractional level of the pyramid applied at each pixel to match the foveal resolution in the formula applied at stage 4 (above). If the maximum frequency of the display device (maxfreq) is equal to the maximum visible frequency (eyefreq), the pyramid level (pyrlevel) is 1. If maxfreq is lower, pyrlevel will be greater than one (that is, at a higher level in the pyramid).

```
pyrlevel = maxfreq ./ eyefreq;
```

6. The pyramid level (pyrlevel) is truncated to conform to the computed levels of the multi-resolution pyramid:

```
pyrlevel = max(1,min(levels,pyrlevel));
```

7. For each of the colour panes in the image frame data, the Gaussian Pyramid [pyr, indices] is constructed and a 3-D matrix (blurtree) is constructed in which each pyramid level is up-blurred by an appropriate factor so that the dimensions of the pyramid level fit those of the original image. This simplifies the interpolation process which completes the formation of the foveated image (outimage):

```
[pyr,indices] = buildGpyr(img,levels);
blurtree = zeros(img_size(1),img_size(2),levels);
for n=1:levels
        blurtree(:,:,n) = ...
                imcrop(upBlur(show, n-1),[1 1 img_size(2)-1 img_size(1)-1]);
outimage(:,:,color_idx) = interp3(blurtree,xi,yi,pyrlevel, '*linear');
```

The degree of foveation is controlled by varying the size of the minimum contrast threshold (CT_0). Raising the CT_0 above the psychophysically measured value produces visible degradation (blurring). However, it is argued that there are tasks where some degradation will not reduce user acceptance. This theory was tested in the subjective quality assessment of spatially foveated BSL video by deaf viewers at a range of CT_0 values (Section 9.2).

8.2 Temporal Video Image Foveation Method

The aim of the experimental work in this section was to develop a method of giving coding priority to important temporal image content in the face region of the signer in the BSL video based on a foveation model. This section describes the rationale and the algorithm developed in Matlab (Appendix F) for this novel method of temporal video image content-prioritisation.

High quality video, measured in terms of the temporal resolution requirements of the HVS, is associated with the perception of smooth motion (Section 2.3.3). The visual perception of motion is influenced by the video frame rate and visual effects such as motion blur and motion sharpening (Section 2.4). The trade-off between video image quality and frame rate presents an optimisation challenge for video communication systems designers. The recommended frame rate for sign language video communication (25 frames per second), determined by ITU-T Study Group 16 (1998), places a significant burden on media content delivery within limited bandwidth resources. The recommendation assumes that the viewer requires uniform provision of temporal information across the whole video scene so that regions that display greatest motion (normally arms and hands in sign language video) can be observed with minimum blurring. However, the viewing behaviour of deaf subjects (described in Section 3.4 and tested in Chapter 6) suggests that this temporal resolution requirement may only be necessary at the point of regard (for processing small rapid temporal changes in the face of the signer in the BSL video image). The results of the eye movement tracking experiments in this thesis demonstrated that deaf viewers do not track the movements of the hands during a signed sequence (Section 6.4.3.3). Furthermore, human peripheral vision is specialised for processing gross motion and deaf subjects have enhanced peripheral processing capabilities (Gazzaniga, Ivry &

Mangun, 1998). These findings suggest that uniform temporal resolution may not be necessary for BSL communication and that temporal resolution requirements could be reduced in the peripheral region of the BSL video image, resulting in a degree of image compression without loss of perceived video quality.

The development of a new method of image compression was based on the hypothesis that the temporal visual requirements of deaf people watching BSL video (foveating on the face and observing hand motion in peripheral vision) are reduced with foveal eccentricity. Building on the model of spatial resolution of the HVS, applied in spatial video image foveation (Section 8.1), a new method of temporal resolution filtering was developed; the Foveation-Weighted Temporal Filtering (FWTF) method for BSL video images. The FWTF method was used to produce BSL video sequences with increased temporal filtering away from the foveation point (that is, the face region of the signer in the BSL video).

The same, psychophysically measured HVS parameters (Table 8.1) applied in the spatial video image foveation model are used to create the pyramid levels for the FWTF. An additional Filter Strength (FS) parameter was created to control the degree of temporal resolution at each level of the foveation pyramid. The value of the Filter Strength parameter determines the Gaussian filter kernel values applied at each pyramid level (Chapter 9, Table 9.6) which are used to create multi-resolution temporal filtering across the video image.

The FWTF algorithm is applied to image data in each frame of the video sequence. The main processing stages in the algorithm are the same as the first six stages which describe the creation of the levels of the multi-resolution pyramid for the spatial video image foveation algorithm (Section 8.1). The final stage of the process (Stage 7) in the FWTF algorithm is described (with Matlab code) as follows:

> The FWTF applies a low-pass Gaussian filter to the video image data (i, j). The length (nfilt) of the Gaussian filter kernel (filtkernel), determines the temporal span for the computation of the temporally foveated image data in the algorithm. For example, a filtkernel value of [0.4987 0.2283 0.0219 0.0040] applies the filter over a temporal span of 4 frames. The filter is not applied to the first (nfilt-1) frames or to data where the pyrlevel is 1. The strength of the filter is determined by the

numerical difference between the values in the kernel (filtkernel). For example, [1 0 0 0] is no filter (the default) and [0.1829 0.1708 0.1390 0.0987] is a strong filter. The calculation of the Gaussian filter (filt) is obtained using the Matlab 'fspecial'[16] function. The value of sigma for the fspecial function is computed from the pyramid level (pyrlevel) and the filter strength (FS) parameter. Pyrlevel is the fractional level of the multi-resolution pyramid used at each pixel to match the foveal resolution function in the image foveation algorithm (as previously described in 8.1). The value of pyrlevel is constrained in the algorithm to conform to the levels of the pyramid which have been computed (that is, up to seven layers in the multi-resolution pyramid model defined by Geisler and Perry (1998)). The value of the Constant in this formula was set to 10; determined by experimentation to obtain a suitable spread of filter strength results.

```
sigma = (((pyrlevel(i,j))-1) + (FS/Constant));
filt = fspecial('gaussian', [1 7], sigma);
```

The upper values [4:7] of the [0:7] array of the rotationally symmetric Gaussian low -pass filter (filt), produced by the fspecial function, are used to create the filter kernel (filtkernel) for the FWTF algorithm.

```
filtkernel = filt(4:7);
```

Examples of the filter kernel values in the application of this method in Chapter 9 are given in Table 9.6.

The temporally filtered sample (outsample) is created from the original pixel data in previous frames (within the temporal span, nfilt) and the filter kernel (filtkernel) values:

```
for f = 1:nfilt
   outsample = outsample + [frame pixel data at current frame -  f +1] * filtkernel(f));
```

[16] fspecial('gaussian',hsize,sigma) returns a rotationally symmetric Gaussian low-pass filter of size, hsize, with standard deviation, sigma.

Figure 8.3 illustrates the application of the FWTF (where, nfilt = 4) to video image pixels in the fourth frame (Frame 3) of a video sequence.

The intensity of the temporally filtered sample is normalised and this completes the formation of the temporally foveated image (outimage).

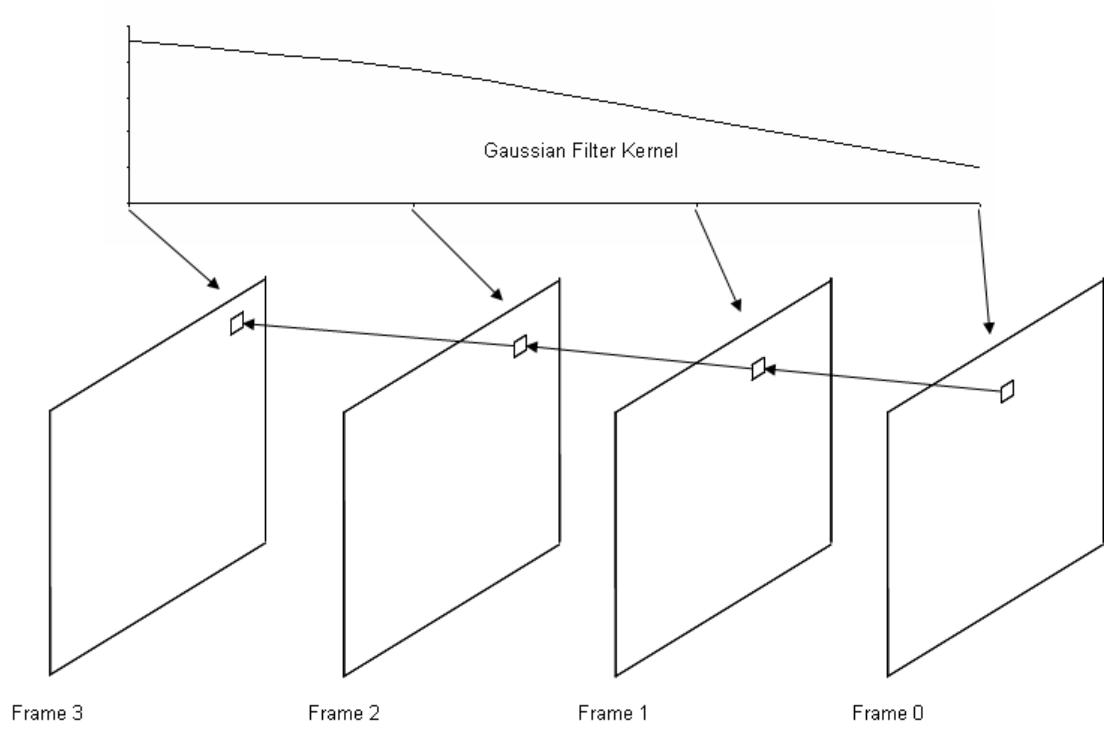


Figure 8.3 Block Diagram of the Application of the Gaussian Filter Kernel in the FWTF Algorithm

Varying the Filter Strength parameter in the FWTF algorithm facilitated the creation of a range of FWTF video sequences from source BSL material for subjective quality testing by deaf viewers (illustrated in Chapter 9, Figure 9.5). This enabled the maximum temporal filter strength to be determined for the perception of acceptable quality video (Section 9.3).

8.3 Summary

The algorithms for spatial and temporal video image foveation described in this chapter produce smoothly foveated images. The degree of foveation is controlled by varying

the minimum contrast threshold (CT_0) parameter in the spatial foveation algorithm and the Filter Strength (FS) parameter in the FWTF algorithm. The application of the algorithms and examples of the outputs are given in Chapter 9 which details experiments conducted to test the impact on BSL communication at a range of degrees of foveation.

Chapter 9: Perceived Quality of Content-Prioritised Video Coding for BSL Communication

Current automatic 'objective', standard subjective and non-standard alternative methods of assessing video quality are inadequate for evaluating quality for the specific task of BSL video communication (discussed in Chapter 5). This chapter describes the development of methods for assessment of perceived quality of BSL video communication by people who are deaf and use BSL as their first/preferred language. The methods are applied in experiments to determine user acceptability of video material which has been pre-processed using the spatial and temporal video image foveation methods developed for this thesis (Chapter 8). Pre-processed video material is evaluated to determine the maximum degree of content-prioritised video image compression which is acceptable to BSL users.

9.1 Subjective Quality Assessment Methods

This section describes the methods developed to facilitate end-user assessment of BSL video sequences. The aim was to develop methods of measuring perceived quality in terms of user satisfaction for the BSL communication task and address the limitations of the current standards-based methods and alternative methods developed in previous research (discussed in Section 5.2). The design criteria for the development of the subjective quality assessment methods for this thesis are specified in Table 9.1. The development of the methods is described with reference to meeting these criteria.

Two subjective quality assessment methods were developed to measure video image quality for BSL communication in this thesis (described in the following sections). In both methods, all communications with subjects were in BSL; no written communication in English language was used in the application of the methods. This was a very important factor in the design of a subjective quality method for use by deaf people communicating in BSL (discussed in Section 3.1) since it minimised user effort (Criterion 8) and avoided any barriers to communication (Criterion 5) which might impact on the results and final analysis.

1	Measure quality in terms of the user task rather than a measure of overall degradation in picture quality (image fidelity).
2	Identify task-specific assessment criteria which can be easily and accurately expressed in BSL for accuracy of reporting user satisfaction and as an indication of ability to perform the BSL communication task (user performance).
3	Ensure reliability and repeatability of results through a structured approach while allowing the opportunity for unstructured free responses (in BSL) as an aid to data analysis.
4	Avoid interference with/disruption to the primary user task.
5	Minimise the influence of external factors (for example, implications of financial cost, language barriers and technology expectations)
6	Minimise the overall length of the experiments with individual deaf subjects.
7	Minimise short-term memory influences.
8	Minimise user effort/skill requirements.
9	Obtain concise user data, efficiently, in a controlled test environment for a wide range of test conditions.
10	Avoid dependence on large sample groups.

Table 9.1 Criteria for the Design and Application of Methods of Subjective Quality Assessment of BSL Video

9.1.1 Task-Specific Category Rating (TSCR) Method

The Task-Specific Category Rating (TSCR) Method was developed to enable assessment of the quality of content-prioritised BSL video with reference to the original (unprocessed) video material (applied in Section 9.2).

The method was based on the structured procedure of the DCR method (Section 5.2.1.2) in ITU Recommendation P.910 (ITU, 1999) to ensure reliability and repeatability of results (Criterion 3) and efficient data gathering for a range of test conditions (Criterion 9). The DCR approach allows direct (double stimulus) comparison between the original (unprocessed) and the processed video in random order of display. The display of the unprocessed and processed video material in a double stimulus test is illustrated in Figure 9.1.

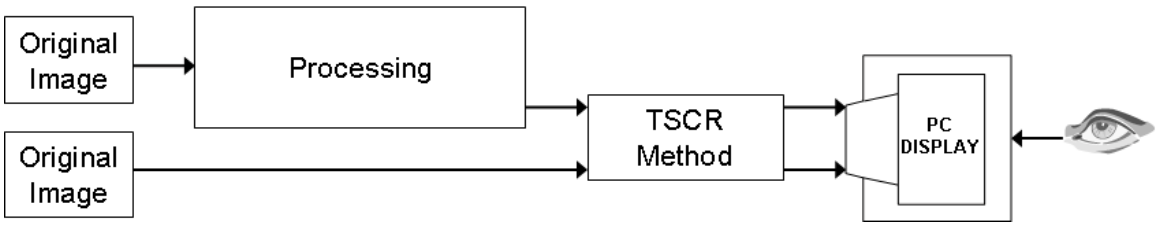


Figure 9.1 Block Diagram of the Task-Specific Category Rating (TSCR) Method

The random order of display minimises financial cost implications associated with preference of the test sequences (Criterion 5) which may be implied in methods which present scalar increase/decrease in test video quality (Section 5.2.3).

The recommended sample size for the DCR method is 4 to 40 subjects (ITU-T Recommendation P.910, 1999). Dependency on large sample groups was not desirable (Criterion 10) for two main reasons. Firstly, the target user represents a small proportion of the general population and secondly, subjective assessment generates a significant amount of qualitative data for each participant, requiring complex analysis. The question of how many users should be included in usability testing was raised by Spiller (2005). He argued that, since usability research is behaviour and task driven, it is mainly qualitative. He maintained that, as the desired end result is a measure of satisfaction in context (rather than statistical validity), a sample size of 10 to 15 participants was sufficient to produce valid qualitative data.

A disadvantage of the structured procedure of the DCR method is the overall duration of the experiment (due to the requirement for repetitions to verify consistency of user response), which conflicts with Criterion 6. In an attempt to address Criterion 6, the duration of individual test sequences and voting periods were kept to the ITU recommended time pattern (Figure 5.2). This also ensured that Criterion 7 could be met since video sequence durations of approximately ten seconds were within short-term memory limits. Continuous rating scales (including the use of a slider) and physiological measures of user cost (discussed in Section 5.2.3) were excluded since it was felt that these methods would interfere with the primary task and therefore would not meet Criterion 4.

The structured procedure of the DCR method was adopted in the TSCR method primarily to obtain data in a controlled test environment for a number of test conditions and to ensure rigour and repeatability of results. However, the limitations of the DCR method for obtaining subjective feedback from subjects in relation to a specific visual task (discussed in Section 5.2.2) provided the rationale for development of a new rating scale.

The new TSCR scale was designed to facilitate assessment of perceived quality for the task rather than overall picture fidelity (Criterion 1). This resulted in the creation of five criterion-referenced rating scale descriptors (Section 9.2.4, Table 9.3). Task performance metrics were not applied since task performance was inferred in the rating descriptors and, according to Dugénie *et al* (2002), this approach would require a large sample group (which conflicts with Criterion 10). An opportunity to gain more qualitative data, which confirmed that the user had performed the task and was satisfied with the video quality for the task, was afforded by permitting free responses at the end of the session with each individual subject (which also partially addressed Criterion 3). The TSCR descriptors were specified in consultation with a qualified BSL Interpreter to ensure that they could be easily translated into BSL and that a clear distinction between each of the points on the rating scale could be made by the assessor (Criterion 2). The TSCR method was also designed to avoid interference with the primary task (Criterion 4) and reduce the effort required by the assessor to make a decision on quality in relation to the task (Criterion 8).

The application of the TSCR method is described and reviewed in Section 9.2.

9.1.2 Binary Acceptability Method

The Binary Acceptability Method was developed to enable end-user assessment of the quality of processed and unprocessed BSL video without direct reference to the original video material (applied in Section 9.3 and in Chapter 10).

The development of the Binary Acceptability Method was based on the 'method of limits' approach by McCarthy, Sasse and Miras (2004), discussed in 5.2.3, and on the experiences of the participants using the TSCR method (reviewed in Section 11.2.1).

The TSCR method allows direct comparison of the test material with reference to the source material in terms of perceived quality for the task in a double stimulus test. The Binary Acceptability Method is a single stimulus approach with decision making on a binary scale; 'acceptable' or 'not acceptable' for the BSL video communication task (Figure 9.2). It addresses the design criteria satisfied by the TSCR method and also reduces the time and effort required by participants (Criteria 6 and 8). The video stimuli are displayed in order of scalar increase/decrease in test video quality (as appropriate) but there are no financial cost implications (Criterion 5) since each test sequence is evaluated independently. Application of the method involves display of each single video stimulus (of up to ten seconds in duration) and voting 'acceptable' or 'unacceptable' (communicated in BSL) for the task in a ten second period after each video sequence (Figure 5.1).

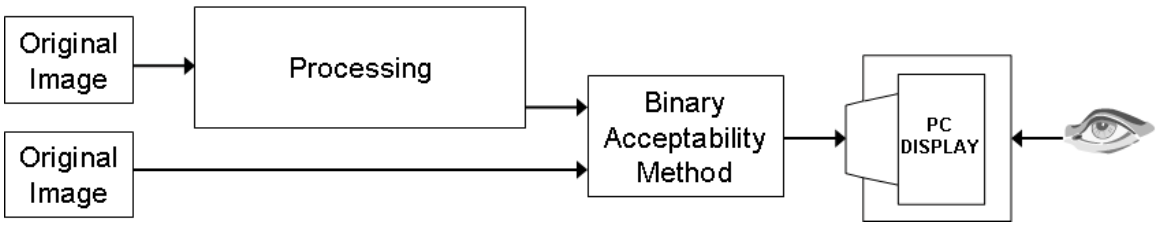


Figure 9.2 Block Diagram of the Binary Acceptability Method of Subjective Testing

This method was developed for quality assessment of discrete video clips (rather than for continuous assessment of quality during a variable quality sequence) to overcome the difficulties of synchronising the user responses with the changing image quality during the sequence (discussed in 5.2.3).

Detail of the procedures and application of the Binary Acceptability Method of subjective testing is given in the description of experiments in Section 9.3 of this chapter and in Chapter 10.

9.2 Perceived Quality of Spatial Video Image Foveation

This section describes an experiment to determine user satisfaction with video material which has been pre-processed using the spatial video image foveation method of content-prioritised coding developed for this thesis (Section 8.1). The perceived quality of pre-processed video material (compared with the original, unprocessed material) is

evaluated using the TSCR method (Section 9.1.1). Analysis of the results determines the maximum degree of spatial video image compression which is acceptable to BSL users.

9.2.1 Experimental Design and Rationale

The objective of this experiment was to evaluate the performance of the spatial video image foveation method and identify the maximum degree of spatial image foveation for perception of acceptable quality for video communication of BSL.

Test BSL video sequences, with a range of degrees of foveation (blurring) from the centre of the face of the signer, were created by varying the minimum Contrast Threshold (CT_0) in the spatial video image foveation algorithm (Section 8.1). A range of test BSL video sequences was produced which included no foveation, foveation at the normal psychophysically measured threshold value for the HVS (Table 8.1) and a series of increasingly foveated sequences with CT_0 values above this threshold. The aim of this approach was to determine the maximum CT_0 value (and thus the highest level of video compression) which would produce acceptable video quality for BSL users.

The limitations of standard subjective quality assessment methods and alternative methods were reviewed and a Task-Specific Category Rating (TSCR) method was developed (Section 9.1.1) which used criterion-referenced rating scale descriptors (Table 9.3) appropriate to the BSL communication task. This method was designed to meet the specific methodological criteria for subjective quality evaluation of BSL video material by deaf subjects (Table 9.1).

The experimental procedure was tested on a sample of hearing subjects before it was conducted with deaf participants. As the hearing subjects had no knowledge of BSL, and thus were not able to undertake the visual task, the results were not analysed or used for a control comparison. No changes were made to the testing regime as a result of this trial as it was clear that the design was appropriate for the experiment.

9.2.2 Subjects

Subjective quality assessment experiments were conducted with six profoundly deaf-from-birth volunteers who use BSL as their first language. Four of the subjects were male (Subjects 1, 4, 5 and 6) and two of the subjects were female (Subjects 2 and 3). The subjects were aged between 46 and 66 years and had normal or corrected-to-normal visual acuity. All communications with subjects were in BSL, aided by a local Interpreter.

9.2.3 Materials and Apparatus

The sign language video material for the experiment was captured at 25 frames per second on a SonyVX200E Digital Video camera, under controlled artificial lighting in the University video recording studio, using one local profoundly deaf volunteer. The signer used facial expression, lip/mouth shapes, gestures, finger-spelling and body movement around the scene which had a plain background. She related short stories from her own experience using her own natural style and expression of signing. Short test BSL video sequences (Table 9.2) were created to include a wide range of sign language movements, expressions and gestures (including finger spelling). The sequences were short (7 to 11 seconds) to minimise the overall time of the experiment for each subject and to ensure that the entire test sequence was given equal weighting in the evaluation within short-term memory limits. In addition, five different video clips were created for training the participants prior to the main experiment.

The video material was pre-processed using the spatial video image foveation technique developed for this thesis (Section 8.1). The test BSL video sequences (except Video 1, which was not foveated) were foveated by stepping through each of the source clips, frame-by-frame[17], marking the central point for foveation (in this case the tip of the nose of the signer which corresponded with the central point of the face) and degrading the spatial quality from the foveation point according to the minimum

[17] Automatic detection and tracking of the foveation point was not implemented at this stage as the objective was to test the hypothesis based on exact knowledge of the foveation point (Section 8.3).

Contrast Threshold (CT_0) and viewing distance set for each video sequence. The CT_0 and thus the degree of foveation blurring ranged from zero (none) to 0.2 (high). The other parameters set in the foveation algorithm remained constant (Table 8.1). The output of the foveation algorithm was a video sequence which had a smooth reduction in spatial resolution from the point of foveation to the edge of the video images.

Video Number	Video Name (Duration)	Minimum Contrast Threshold (CT_0)	English Translation of BSL Content
1	Lisa Bus Stop (10.20 seconds)	0.00	"Standing at the bus stop, people could see I was deaf. They could see my hearing dog's jacket but they still talked. I didn't know what they were talking about."
2	Lisa Family (10.16 seconds)	0.0156	"I have one brother and one sister. I am older than them. My brother is divorced. He has two children."
3 (& 10)	Lisa Introduction (10.18 seconds)	0.03 (& 0.0156en)	"Hello, my name is Lisa. My dog's name is Bran. He is a hearing dog for the deaf. He helps me."
4	Lisa Hobbies (7.22 seconds)	0.05	"My hobbies, I love cooking, swimming and tap dancing."
5	Lisa Holiday (10.10 seconds)	0.075	"I went on holiday to Spain last year. I had a good time. The weather was very warm."
6	Lisa School (8.03 seconds)	0.10	"I went to Aberdeen School for the Deaf. As I was growing up I used signs and learned oral communication."
7	Lisa Worlds (7.02 seconds)	0.13	"When I was at school there was a hearing world and a deaf world. Now I have both worlds."
8	Lisa Deaf (9.05 seconds)	0.16	"When you meet a deaf person you think deaf people are the same. They are not; there are different levels of deafness."
9	Lisa Television (9.01 seconds)	0.20	"When I watch TV, I look at the subtitles or the signer at the bottom right and I look back and to the TV picture."

Table 9.2 Video Material for Subjective Quality Testing

An additional sequence with 'enhanced' foveation (Test BSL Video Sequence 10) was created, from the source for Video Sequence 3, by setting the CT_0 to the normal psychophysically measured value for the HVS (CT_0 = 0.0156) and increasing the value

of alpha to 0.212 and decreasing the value of e_2 to 1.15. This 'enhanced' effect was described by Geisler and Perry (1998) and was implemented in this thesis to test the response to an extreme degree of foveation (CT_0 = 0.0156en).

The video clips for the training session, conducted before the main experiment, were created with CT_0 values of 0.0156, 0.05, 0.075, 0.10 and 0.20.

9.2.4 Procedure

The experiment was conducted with individual subjects positioned at a comfortable viewing distance from a seventeen-inch monitor (true colour, 32 bit display) connected to a Dell Pentium IV PC with PCI Video Capture Card installed. The test BSL video sequences were presented full screen one at a time, grouped in pairs. The first video in the pair was the reference (source) video sequence and the second sequence was the source video which had been foveated according to the CT_0 value set in the foveation algorithm. A plain mid-grey coloured screen was presented for two seconds between each sequence in the pair and for ten seconds between each pair of sequences. The subjects were asked to rate the quality of the second (foveated) video sequence compared to the first (original unprocessed reference) sequence during the ten-second period (voting time) between pairs of sequence. The experimental set-up is illustrated in Figure 9.3.

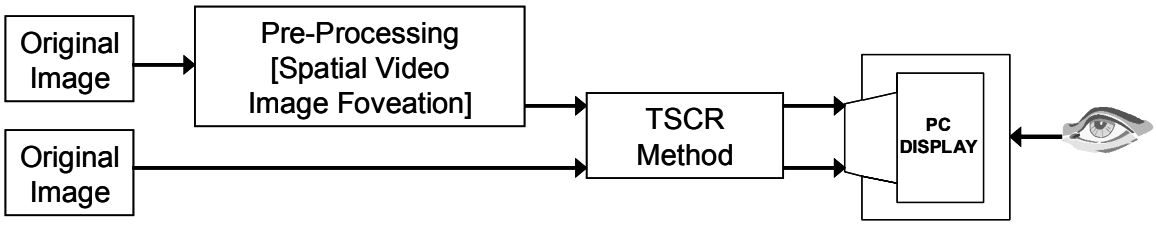


Figure 9.3 Block Diagram of the Subjective Quality Evaluation of Original (Unprocessed) and Pre-Processed (Spatially Foveated) BSL Video Material

Subjects used the TSCR scale, which was developed for this thesis to provide criteria for subjective assessment of perceived quality in terms of the BSL communication task (discussed in Section 9.1.1). The rating scale descriptors were translated into BSL by

the local Interpreter. The five-point TSCR rating scale, with English translations of the descriptors, is given in Table 9.3.

Prior to the main experiment, a training session consisting of six video pairs was conducted with each participant. The purpose of the training session was to familiarise the subject with the procedure and the rating scale.

Rating	Description
5	Imperceptible difference
4	Perceptible difference but not annoying, the sign language was clear
3	Sign language is slightly unclear, one or two signs were not clear but the story was understood
2	Annoying, sign language was not clear making it difficult to understand the story
1	Very annoying, sign language was obscured and the story could not be understood

Table 9.3 Criterion-Referenced Five-Point Task-Specific Category Rating (TSCR) Scale for Quality Assessment of BSL Video Communication

The results for the training session were recorded (so that the experimental conditions were exactly the same) but not included in the analysis. The main experiment consisted of six sets of five pairs of video sequences (Tests A to F, Table 9.4) with short rest breaks between each set of video pairs. The clips were presented in random order (not according to the degree of foveation) and each video was included three times in a different order during the experiment. The purpose of the repetition was to allow reliability of scoring by individual subjects to be checked.

9.2.5 Results

Sign language conversations with the subjects after the experiment demonstrated understanding of, and interest in, the content of the video clips used. The TSCR subjective quality ratings (described in Table 9.3) given by each of the six subjects during six sets of five video sequence pairs (Tests A to F) are given in Table 9.4. This

table also provides the Mean Opinion Score (MOS) and Standard Deviation (SD) for each viewing of the clips in the experiment.

A high MOS (rating above 4) was obtained at each level of foveation in the experiment (Figure 9.4). Foveated sequences created with CT_0 values up to and including 0.13 received a higher MOS than the original (unprocessed) reference sequence. This was specifically reported in feedback from Subject 4, who described the picture quality in two of the tests (at CT_0 set to 0.03 and 0.075) as being better than the original for sign language comprehension. At CT_0 values greater than 0.13 (including the very high degree of blurring in the 'enhanced' foveation sequence), the MOS was lower than the MOS for the reference sequence but remained high at over 4.2.

The ratings awarded by each subject, recorded at each of three instances of viewing the clips, and the MOS and Standard Deviation at each level of foveation, are given in Table 9.5.

The reliability of the scoring by the individual subjects was checked with reference to the Standard Deviation for each test sequence. Three of the subjects (Subjects 1, 2 and 4) showed some slight inconsistencies (with standard deviation of 1.0 or greater) in their scoring (emboldened figures in grey cells of Table 9.5). Subject 4 rated one of the test sequences with different scores but this was only at the very highest level of foveation (the 'enhanced' foveated sequence). The lowest score given by this subject in this case was for the second display of this sequence, that is a score of 3 (indicating that one or two signs were not clear) although he rated the same sequence with a score of 5 (imperceptible difference) on the previous and subsequent showings. Subject 2 displayed some uncertainty in the scoring of the test sequences foveated at CT_0 values of 0.16 and 0.075. At CT_0 set to 0.16, she rated the sequence with a low score of 2 (that is, the BSL was not clear making it difficult to understand the story) on the first showing but with higher scores on subsequent showings during the experiment. There were only two other instances of a score of 2 being used to rate the test sequences (both by Subject 1, for the sequence foveated at CT_0 set to 0.16 and the 'enhanced' foveation clip). In each of these cases, the same clips were awarded a higher score (4 or 5) at a different point in the experiment. There were no cases of a score of 1 being awarded in the experiment. Subject 2 rated the sequence foveated at CT_0 set to 0.075 at a score of 3 for the first showing and a high score of 5 at

subsequent showings. The responses from Subject 1 had the most inconsistencies in the scores awarded for repetitions of the test sequences (at CT_0 set to 0.05, 0.075, 0.16, 0.2 and for the 'enhanced' foveated sequence).

The increase in the spread of the results at higher levels of foveation (above CT_0 set to 0.16) was affected by the responses of Subjects 1 and 2. However, it was felt that the reliability of the scoring for the sample group as a whole was generally good and that most of the minor inconsistencies occurred at the higher foveation levels. The Standard Deviation of the scores at each foveation level for the sample group was low up to CT_0 set to 0.16.

An 'optimum' level of foveation, that is a point beyond which the acceptability level dropped below a threshold value, was not evident in the results. However, the maximum foveation level (CT_0) which produced the highest average MOS (4.9) with the lowest standard deviation (0.2) was found to be 0.1 (Table 9.5). This CT_0 value was used to create encoded test video sequences for further subjective quality assessments at fixed bit rates (Chapter 10).

Test A		TSCR Score per Subject							
Video	CTo	**1**	**2**	**3**	**4**	**5**	**6**	MOS	SD
10	0.0156en	3	5	5	5	3	5	4.3	1.0
2	0.0156	5	5	5	5	5	5	5.0	0.0
4	0.0500	4	5	5	5	5	5	4.8	0.4
5	0.0750	5	3	5	5	5	5	4.7	0.8
6	0.1000	5	5	5	5	4	5	4.8	0.4

Test B		TSCR Score per Subject							
Video	CTo	**1**	**2**	**3**	**4**	**5**	**6**	MOS	SD
1	0.0000	5	4	5	5	5	5	4.8	0.4
3	0.0300	5	5	5	5	5	5	5.0	0.0
9	0.2000	3	5	5	4	4	5	4.3	0.8
7	0.1300	5	5	5	5	5	5	5.0	0.0
8	0.1600	2	5	5	5	5	5	4.5	1.2

Test C		TSCR Score per Subject							
Video	CTo	**1**	**2**	**3**	**4**	**5**	**6**	MOS	SD
3	0.0300	5	5	5	5	5	5	5.0	0.0
1	0.0000	4	4	5	4	5	4	4.3	0.5
2	0.0156	5	5	5	5	5	4	4.8	0.4
4	0.0500	3	5	4	5	5	5	4.5	0.8
7	0.1300	4	5	4	5	5	5	4.7	0.5

Test D		TSCR Score per Subject							
Video	CTo	**1**	**2**	**3**	**4**	**5**	**6**	MOS	SD
5	0.0750	5	5	4	5	5	5	4.8	0.4
8	0.1600	3	5	4	5	5	5	4.5	0.8
6	0.1000	5	5	5	5	5	5	5.0	0.0
2	0.0156	5	4	4	5	5	5	4.7	0.5
1	0.0000	4	5	4	5	5	5	4.7	0.5

Test E		TSCR Score per Subject							
Video	CTo	**1**	**2**	**3**	**4**	**5**	**6**	MOS	SD
9	0.2000	5	4	4	3	5	5	4.3	0.8
10	0.0156en	2	5	4	3	4	5	3.8	1.2
5	0.0750	3	5	5	5	5	4	4.5	0.8
4	0.0500	5	5	5	5	5	5	5.0	0.0
8	0.1600	5	5	5	5	5	5	5.0	0.0

Test F		TSCR Score per Subject							
Video	CTo	**1**	**2**	**3**	**4**	**5**	**6**	MOS	SD
7	0.1300	5	5	5	5	4	5	4.8	0.4
9	0.0200	5	5	5	5	4	5	4.8	0.4
6	0.1000	5	5	5	5	5	5	5.0	0.0
3	0.0300	5	5	5	5	5	5	5.0	0.0
10	0.0156en	5	5	4	5	4	5	4.7	0.5

Table 9.4 Subjective Quality Scores (Raw Data) for Tests A-F

CTo	Subject 1					Subject 2					Subject 3					Subject 4					Subject 5					Subject 6					Average	
	1	2	3	MOS	SD	1	2	3	MOS	SD	1	2	3	MOS	SD	1	2	3	MOS	SD	1	2	3	MOS	SD	1	2	3	MOS	SD	MOS	SD
0.0000	5	4	4	4.3	0.6	4	4	5	4.3	0.6	5	5	4	4.7	0.6	5	4	5	4.7	0.6	5	5	5	5.0	0.0	5	4	5	4.7	0.6	4.6	0.5
0.0156	5	5	5	5.0	0.0	5	5	4	4.7	0.6	5	5	4	4.7	0.6	5	5	5	5.0	0.0	5	5	5	5.0	0.0	5	4	5	4.7	0.6	4.8	0.4
0.0300	5	5	5	5.0	0.0	5	5	5	5.0	0.0	5	5	5	5.0	0.0	5	5	5	5.0	0.0	5	5	4	4.7	0.6	5	5	5	5.0	0.0	4.9	0.2
0.0500	4	3	5	4.0	1.0	5	5	5	5.0	0.0	5	4	5	4.7	0.6	5	5	5	5.0	0.0	5	5	5	5.0	0.0	5	5	5	5.0	0.0	4.8	0.5
0.0750	5	5	3	4.3	1.2	3	5	5	4.3	1.2	5	4	5	4.7	0.6	5	5	5	5.0	0.0	5	5	5	5.0	0.0	5	5	4	4.7	0.6	4.7	0.7
0.1000	5	5	5	5.0	0.0	5	5	5	5.0	0.0	5	5	5	5.0	0.0	5	5	5	5.0	0.0	4	5	5	4.7	0.6	5	5	5	5.0	0.0	4.9	0.2
0.1300	5	4	5	4.7	0.6	5	5	5	5.0	0.0	5	4	5	4.7	0.6	5	5	5	5.0	0.0	5	5	4	4.7	0.6	5	5	5	5.0	0.0	4.8	0.4
0.1600	2	3	5	3.3	1.5	2	4	3	3.0	1.0	5	4	5	4.7	0.6	5	5	5	5.0	0.0	5	5	5	5.0	0.0	5	5	5	5.0	0.0	4.3	1.1
0.2000	3	5	4	4.0	1.0	5	4	5	4.5	0.5	5	4	5	4.5	0.5	4	3	4	3.5	0.5	4	5	5	4.5	0.5	5	5	5	5.0	0.0	4.3	0.7
0.0156en	3	2	5	3.3	1.5	5	5	5	5.0	0.0	5	4	4	4.3	0.6	5	3	5	4.3	1.2	3	4	4	3.7	0.6	5	5	5	5.0	0.0	4.3	1.0

Table 9.5 TSCR Scores Awarded by Subjects 1 to 6 on Three Occasions of Viewing each Test Sequence at Different Levels of Foveation (CT_0) and the MOS (Mean Opinion Score) and SD (Standard Deviation) for each Subject and each Test

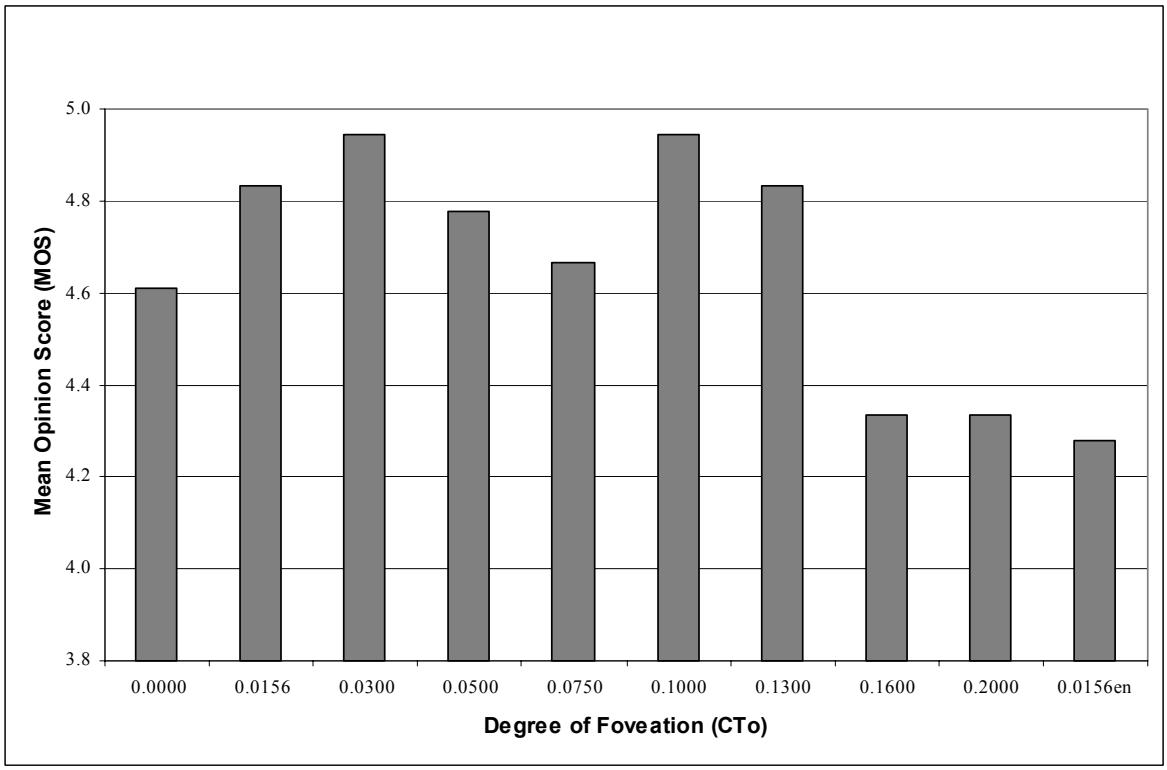


Figure 9.4 Mean Opinion Score (on TSCR Scale) at each Degree of Foveation (CT_0)

9.2.6 Discussion

The aim of this experiment was to evaluate the spatial video image foveation technique and to determine the maximum degree of image foveation (and thus the highest level of video image compression) which would produce acceptable perceived video quality for BSL users in subjective quality evaluation tests. The TSCR method of measuring perceived quality of the test BSL video sequences is evaluated for achieving this aim.

The application of the TSCR rating scale was successful in obtaining reliable data on the level of user satisfaction with pre-processed (foveated) video sequences, compared to the unprocessed reference material, within the design constraints identified for the experiment (Table 9.1). The subjective quality assessment method ensured that quality of the pre-processed material was measured in terms of user satisfaction for the task (rather than overall image fidelity) and that the criteria for assessment of quality were

clear. The method of obtaining feedback ensured that there was no interference with the primary task, no need for any special user skills and that the influence of any external factors (such as language barriers or perceptions of the 'cost' of quality) was eliminated. The short duration of each test video sequence ensured that the user was able to make comparisons between the source and pre-processed video within a comfortable short-term memory span. This also enabled the experiment for each subject to be completed within 25 minutes (including time for additional feedback in BSL through the Interpreter). The short rest breaks between sets of video pairs in the experiment minimised the tedium for participants given the number of comparisons at a range of foveation levels, the number of repetitions at different points in the test (to ensure reliability of results) and the level of visual concentration required. However, it was felt that the time and effort required from each subject (and the BSL Interpreter) could be reduced to improve the experimental conditions for participants (discussed in Section 11.2.1) and this was addressed in subsequent tests (Section 9.3 and Chapter 10).

The spatial video image foveation technique was successful in producing 'smooth' blurring which increased away from the face of the signer in the video. The results demonstrated that deaf viewers accepted pre-processed video at high levels of foveation as long as the BSL communication was clear. This verified the observations of Siple (1978) and validated the findings of the eye movement tracking experiments in this thesis (Chapter 6) since high quality was associated with high resolution of the face in the foveated video images, suggesting that this was the most important region of the video image for BSL communication. The blurring produced by the foveation algorithm in the peripheral region of the image was not perceived by the subjects foveating on the 'clear' face region.

An interesting observation from the results was that the MOS (Mean Opinion Score) for the test video sequences foveated at CT_0 values up to and including 0.13 was higher than the MOS for the original (unprocessed source) video sequence (Figure 9.4). One of the subjects (Subject 4) specifically commented that two of the test video sequences (foveated at CT_0 values of 0.03 and 0.075) were 'better' than the original for sign language communication. These results were recorded as '5' on the TSCR scale but it could be argued that a rating of '6' could have been included to account for quality

which was perceived as being better than the original (unprocessed) sequence. It was not expected that foveation would improve the perception of video image quality (compared to the unprocessed source material) during the design of this experiment and so a rating of '6' was not included. However, this feedback was made possible from the comments made by the participants after the main experiment was complete and so was an important contribution to the final analysis.

It was concluded that spatial video image foveation can be applied to enhance the perceived quality of BSL video content (by giving priority to the face of the signer) and that this is evident up to and including a minimum Contrast Threshold value of 0.1 (that is, the maximum level of foveation producing the highest MOS with the lowest SD). This may be due to the removal of redundant background information which can not be processed in peripheral vision when visual attention is on the facial region of the signer in the video in this complex task.

The implications of these results are important in the design of video communication systems for deaf people with specific visual information requirements. This experiment enabled the determination of the value of CT_0 as an indicator of the critical limit of spatial resolution requirements for BSL video communication. An additional experiment was conducted (Section 9.3) to determine whether there was a similar limit for temporal resolution requirements based on a temporal model of video image foveation, FWTF described in Section 8.2. The critical CT_0 value of 0.1 identified in this experiment was applied in an experiment to test user acceptability of foveated BSL video sequences encoded at a range of fixed bit rates in this thesis (Chapter 10). However, it was acknowledged that, as there is natural variation in the visual requirements of the wider deaf community, a video communication system designed to meet their individual needs must be adaptable to a range of visual thresholds.

9.3 Perceived Quality of Temporal Video Image Foveation

This section describes an experiment to determine user satisfaction with video material which has been pre-processed using the temporal video image foveation (FWTF) method of content-prioritised coding developed for this thesis (described in Section 8.2). The perceived quality of the pre-processed video material is evaluated using the Binary

Acceptability Method (Section 9.1.2). Analysis of the results determines the maximum degree of temporal video image compression which is acceptable to BSL users.

9.3.1 Experimental Design and Rationale

The objective of this experiment was to determine user satisfaction with BSL video which had been pre-processed at a range of FWTF Filter Strengths and to determine the maximum degree of FWTF (that is, the maximum temporal Filter Strength) which would provide acceptable quality for BSL video communication.

User acceptability of the video (which was pre-processed using the FWTF method) was determined using a Binary Acceptability Method (Section 9.1.2) based on the approach by McCarthy, Sasse and Miras (2004). The rationale for development of this approach was that it would meet the experimental criteria identified in Table 9.1 with particular emphasis on keeping the duration of the experiment to a minimum based on the experience of the subjective quality evaluation of spatially foveated BSL video images (Section 9.2).

The application of this method involved the display of a single stimulus rather than comparison with the source material at each FWTF Filter Strength to reduce the time and effort required from the participants. The sequences were displayed in order of increasing temporal Filter Strength and not in the random order which was appropriate for the TSCR method applied in Section 9.2. The rationale was that this would enable the subject to determine the point at which video became unacceptable with reference to the previously acceptable sequences, thus easing decision-making for the subject and reducing the number of repetitions of video sequences which were required in the previous subjective quality evaluation experiments (Section 9.2). However, the subject could request repetition of a sequence at any point in the experiment if further reference points were required.

The experimental method and procedure were tested on a sample of six hearing subjects (who were given specific viewing instructions) before the test was conducted with six deaf participants. As the hearing subjects had no knowledge of BSL and were thus not able to evaluate video quality with respect to the visual/cognitive task, the results were not included in the testing of the hypothesis. However, the results were

used to determine the suitability of the method and they were also used for a controlled comparison of the responses for different viewing patterns (Section 9.2.3). The rationale for giving different viewing instructions to the hearing subjects was to determine whether this had an effect on the results and if any comparisons could be made with the viewing behaviour of the deaf subjects. No changes were made to the testing regime as a result of this trial as it was clear that the design was appropriate for the experiment.

9.3.2 Subjects

The subjective quality assessment experiments for hypothesis testing in this experiment were conducted with the same six profoundly deaf-from-birth volunteers and local BSL Interpreter who participated in the spatial image foveation experiment described in Section 9.2. Six hearing subjects with no knowledge of BSL participated in the initial control experiment.

9.3.3 Materials and Apparatus

The sign language video material for this experiment was created from Test BSL Video Sequence 9 (Table 9.2). This sequence (262 frames, 9.01 seconds) was selected for the experiment as it includes wide and rapid motion of the arms and hands away from the face of the signer (in the peripheral region of the image).

The test BSL video sequence was processed by stepping through each frame and marking the central point for foveation (in this case the tip of the nose of the signer), as described for the spatial image foveation method. The (x, y) coordinates of each foveation point in the sequence were stored in a Matlab data file for input to the FWTF algorithm.

The temporal resolution of each video frame was degraded from the foveation point according to the Filter Strength (FS) variable parameter in the FWTF algorithm. The other parameters set in the FWTF algorithm remained constant; viewing distance = 0.305, spatial frequency decay constant (alpha) = 0.106 and half-resolution retinal eccentricity (e_2) = 2.3 (as described in Table 8.1).

Six test BSL sequences were created, the original unprocessed sequence (with no FWTF applied) and five further sequences created by setting the Filter Strength (FS) parameter in the FWTF algorithm (described in Section 8.2) to 2 (weak temporal filter), 4, 6, 8 and 10 (strong temporal filter). The Gaussian filter kernel values at each level of the foveation pyramid, for each of the five Filter Strength values set in the algorithm, are detailed in Table 9.6.

The output of the FWTF algorithm was a set of video sequences which had reduced temporal resolution away from the point of foveation (with the 'clear' foveated area set as the face of the signer using the viewing distance parameter) to the edge of the video image. Sample frames (Figure 9.5) from the test material illustrate the effect of increasing the temporal Filter Strength on the rapid motion of the right arm of the signer at the start of the sequence. The original unprocessed image from the sequence (Figure 9.5(a)) is blurred (at a frame rate of 25 frames per second) but the blurred edges of the hand can be clearly seen. At a Filter Strength set to 10 (Figure 9.5(f)) the shape of the hand is not defined and only the motion trail is visible in this frame.

FS=2		Filter Kernel Values			
Pyrlevel	*Sigma*	**1**	**2**	**3**	**4**
1.0	*0.2*	1.0000	0.0000	0.0000	0.0000
1.2	*0.4*	0.9192	0.0404	0.0000	0.0000
1.4	*0.6*	0.6638	0.1655	0.0026	0.0000
1.6	*0.8*	0.4987	0.2283	0.0219	0.0040
1.8	*1.0*	0.3991	0.2420	0.0540	0.0044
2.0	*1.2*	0.3333	0.2356	0.0831	0.0146
2.2	*1.4*	0.2880	0.2232	0.1038	0.0290
2.4	*1.6*	0.2560	0.2106	0.1172	0.0441
2.6	*1.8*	0.2330	0.1997	0.1257	0.0581
2.7	*1.9*	0.2239	0.1950	0.1287	0.0644

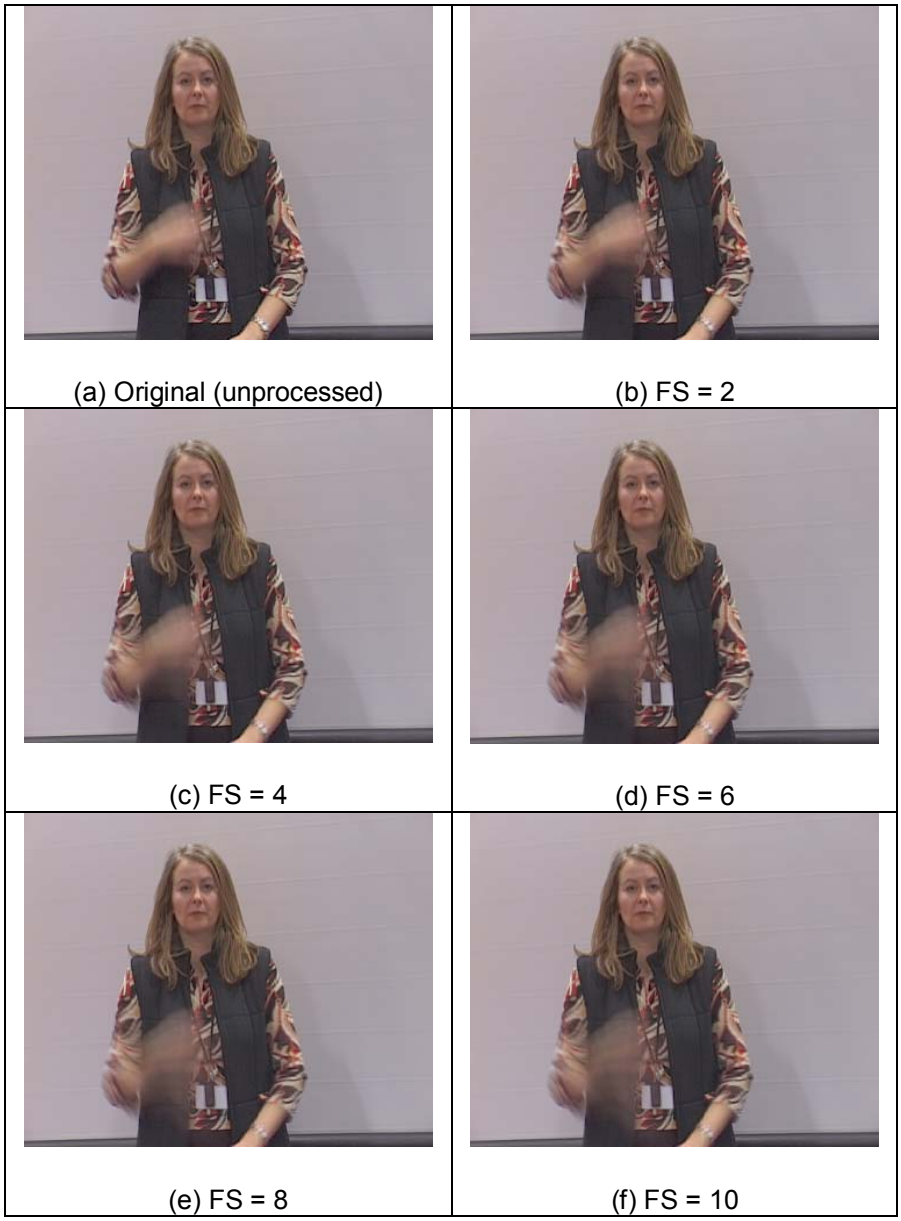
FS=4		Filter Kernel Values			
Pyrlevel	*Sigma*	**1**	**2**	**3**	**4**
1.0	*0.4*	0.9192	0.0404	0.0000	0.0000
1.2	*0.6*	0.6638	0.1655	0.0026	0.0000
1.4	*0.8*	0.4987	0.2283	0.0219	0.0040
1.6	*1.0*	0.3991	0.2420	0.0540	0.0044
1.8	*1.2*	0.3333	0.2356	0.0831	0.0146
2.0	*1.4*	0.2880	0.2232	0.1038	0.0290
2.2	*1.6*	0.2560	0.2106	0.1172	0.0441
2.4	*1.8*	0.2330	0.1997	0.1257	0.0581
2.6	*2.0*	0.2161	0.1907	0.1311	0.0702
2.7	*2.1*	0.2093	0.1869	0.1330	0.0755

FS= 6		Filter Kernel Values			
Pyrlevel	*Sigma*	**1**	**2**	**3**	**4**
1.0	*0.6*	0.6638	0.1655	0.0026	0.0000
1.2	*0.8*	0.4987	0.2283	0.0219	0.0040
1.4	*1.0*	0.3991	0.2420	0.0540	0.0044
1.6	*1.2*	0.3333	0.2356	0.0831	0.0146
1.8	*1.4*	0.2880	0.2232	0.1038	0.0290
2.0	*1.6*	0.2560	0.2106	0.1172	0.0441
2.2	*1.8*	0.2330	0.1997	0.1257	0.0581
2.4	*2.0*	0.2161	0.1907	0.1311	0.0702
2.6	*2.2*	0.2034	0.1835	0.1346	0.0803
2.7	*2.3*	0.1982	0.1804	0.1358	0.0847

FS= 8		Filter Kernel Values			
Pyrlevel	*Sigma*	**1**	**2**	**3**	**4**
1.0	*0.8*	0.4987	0.2283	0.0219	0.0040
1.2	*1.0*	0.3991	0.2420	0.0540	0.0044
1.4	*1.2*	0.3333	0.2356	0.0831	0.0146
1.6	*1.4*	0.2880	0.2232	0.1038	0.0290
1.8	*1.6*	0.2560	0.2106	0.1172	0.0441
2.0	*1.8*	0.2330	0.1997	0.1257	0.0581
2.2	*2.0*	0.2161	0.1907	0.1311	0.0702
2.4	*2.2*	0.2034	0.1835	0.1346	0.0803
2.6	*2.4*	0.1937	0.1776	0.1369	0.0887
2.7	*2.5*	0.1897	0.1751	0.1377	0.0923

FS=10		Filter Kernel Values			
Pyrlevel	*Sigma*	**1**	**2**	**3**	**4**
1.0	*1.0*	0.3991	0.2420	0.0540	0.0044
1.2	*1.2*	0.3333	0.2356	0.0831	0.0146
1.4	*1.4*	0.2880	0.2232	0.1038	0.0290
1.6	*1.6*	0.2560	0.2106	0.1172	0.0441
1.8	*1.8*	0.2330	0.1997	0.1257	0.0581
2.0	*2.0*	0.2161	0.1907	0.1311	0.0702
2.2	*2.2*	0.2034	0.1835	0.1346	0.0803
2.4	*2.4*	0.1937	0.1776	0.1369	0.0887
2.6	*2.6*	0.1861	0.1728	0.1384	0.0956
2.7	*2.7*	0.1829	0.1708	0.1390	0.0987

Table 9.6 FWTF Gaussian Filter Kernel Values at Filter Strength (FS) equal to 2,4,6,8 and 10



(a) Original (unprocessed) (b) FS = 2

(c) FS = 4 (d) FS = 6

(e) FS = 8 (f) FS = 10

Figure 9.5 Original (Unprocessed) and Pre-processed (FWTF at a Range of Filter Strengths (FS)) Images for Test BSL Video Sequence 9 (Frame 20); (a) Original (Unprocessed), (b) FS = 2, (c) FS = 4, (d) FS = 6, (e) FS = 8 and (f) FS = 10

9.3.4 Procedure

The six test BSL video sequences were displayed to each participant on a seventeen inch monitor with true colour, 32 bit display connected to a Dell Pentium IV PC with PCI Video Capture Card installed.

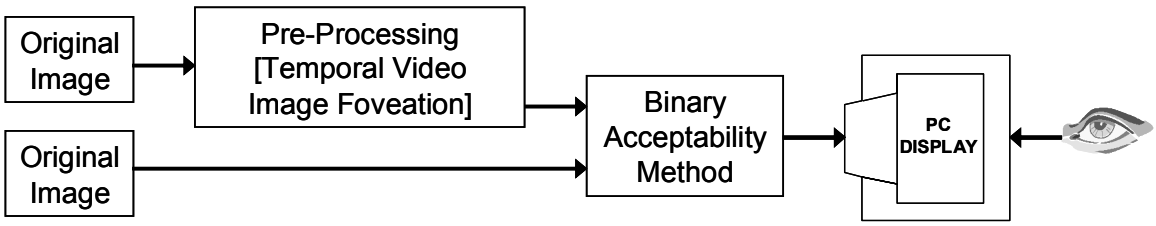


Figure 9.6 Block Diagram of the Subjective Quality Evaluation of Original (Unprocessed) and Pre-Processed (Temporally Foveated) BSL Video Material

The Binary Acceptability Method (Section 9.1.2) of subjective quality assessment procedure was conducted with individual subjects positioned at a comfortable viewing distance from the PC monitor (Figure 9.6). The test BSL video sequences were presented full screen one at a time in increasing order of temporal Filter Strength. Subjects were asked to rate the quality of each clip as either acceptable or unacceptable. In the initial control experiment, hearing subjects were instructed to rate their perception of quality of the FWTF video images for two different views of the video sequences; firstly, after looking at the face of the signer only and secondly, after free viewing of each of the video sequences. In the experiment conducted with deaf subjects, the subjects were instructed to rate their perception of the quality of the FWTF video images once for BSL communication. No instructions were given to deaf subjects on how the video sequences should be viewed.

9.3.5 Results

In the experiment conducted with deaf subjects, 4 of the subjects (Subjects 2, 3, 5 and 6) rated all video sequences as being acceptable and the remaining 2 (Subjects 1 and 4) subjects found the clips to be acceptable up to and including a Filter Strength value of 6. These two subjects reported “haziness” in the test video sequences beyond this

Filter Strength value. None of the subjects requested repetitions of any of the video sequences in the test and all were satisfied with the overall duration of the test.

In the experiment conducted with hearing subjects, the results varied according to the viewing instructions. In part (a) of the experiment (subjects looking at the face of the signer), three of the subjects rated all video clips as being acceptable, two of the subjects rated the video clips as being acceptable up to and including a Filter Strength value of 4 and similarly, one up to and including a Filter Strength of 6. In part (b) of the experiment (free viewing of the video clips), one of the subjects rated all video clips as being acceptable and the remaining five subjects all rated the sequences unacceptable at a Filter Strength of 2 and beyond. They reported a visible trailing or blurring effect in the movements of the signer in the clip at this Filter Strength and beyond. The results are plotted in Figure 9.7 (error bars are included at the 95% confidence interval).

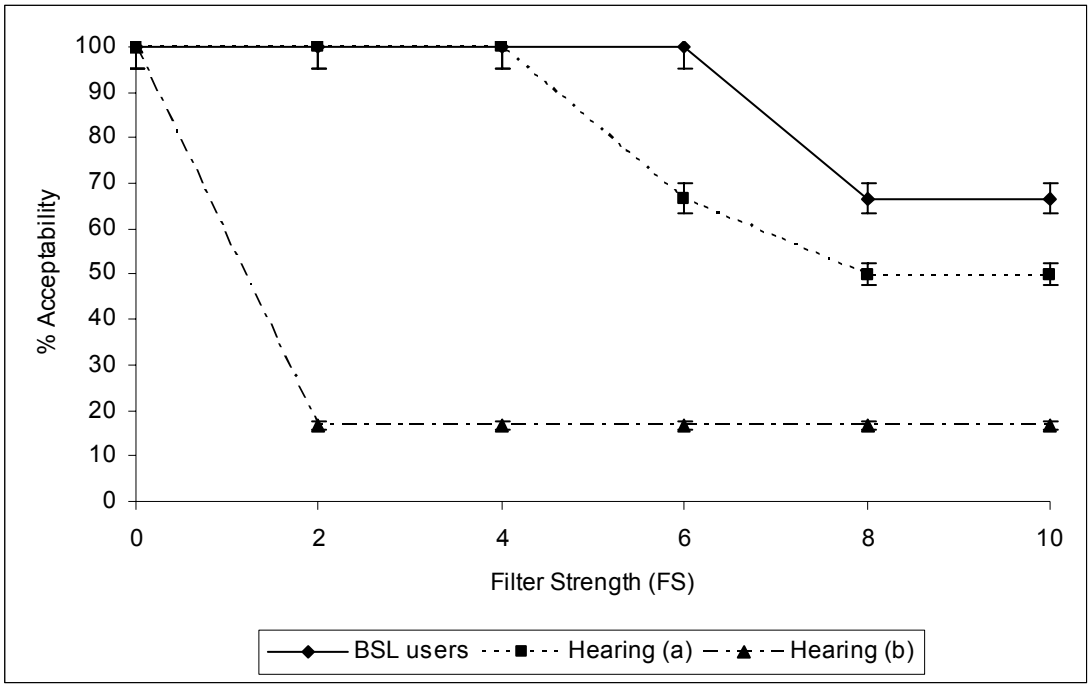


Figure 9.7 Acceptability of FWTF Method for BSL Users, Hearing Group (a) and Hearing Group (b) at Filter Strengths (FS) from 0 to 10

The advantage of plotting the results as percentage acceptability is that the results can be mapped to utility curves.

9.3.6 Discussion

The hypothesis for this experiment was that increased temporal filtering away from the point of regard, using a FWTF, results in bandwidth savings without loss of subjective quality in sign language video. Evidence was obtained to support this hypothesis. The highest FWTF Filter Strength at which subjective quality was 100% acceptable for deaf participants was 6 in the Binary Acceptability test developed for this thesis. The Binary Acceptability Method was successful in obtaining concise data for analysis and in minimising the time and effort required from the participants.

The results for deaf subjects were very similar to those for hearing subjects when they were instructed to look at only the face region of the signer. Assuming that both groups of subjects were actually foveating on the face of the signer in the video, deaf subjects appeared to accept the quality offered at a higher FWTF Filter Strength (that is, lower temporal resolution in the peripheral region), perhaps due to superior concentration on the face of the signer in the sequence and/or enhanced information processing capabilities in peripheral vision for the visual task. The results for free viewing of the video sequences by hearing subjects suggest that gaze around the scene made the blurring effect of the FWTF evident. However, this was not the case for hearing Subject 4 whose gaze was perhaps mostly on the face by choice in this test.

These results demonstrate that image compression (by removal of visually redundant temporal image quality in the peripheral field of view) was achieved while maintaining acceptable perceived quality for BSL video communication.

9.4 Objective Analysis of the Content-Prioritisation Methods

An ‘objective’ analysis of the BSL content-prioritisation methods was conducted to determine the potential coding gain (in terms of coding efficiency) of applying and combining the methods in a standard video coding scheme. The analysis was conducted using Test BSL Video Sequence 9 (described in Table 9.2).

The bandwidth requirements (encoded bit rate) of the original (unprocessed) sequence, the spatially foveated sequence (foveated at $CT_0 = 0.1$), the FWTF sequence (FS = 6) and a combined spatio-temporal foveated filter ($CT_0 = 0.1$ and FS = 6) was investigated.

The original and test BSL video sequences were encoded using the H.264/JVC Joint Model (JM9.2) reference software CODEC. The bit rate and PSNR obtained for each sequence is given in Table 9.7. The bit rate for the original and pre-processed sequences is plotted against the Quantisation Parameter (QP) in Figure 9.8.

Significant savings in bandwidth requirements were obtained for the pre-processed BSL video sequences compared to the source video at the same QP (Table 9.7 and Figure 9.8). The spatial video image foveation method of BSL content-prioritisation (at $CT_0 = 0.1$) achieved a bit rate saving of between 39 and 48% compared to the original unprocessed video sequence. The relative value (in terms of bit rate savings) of the FWTF method (FS = 6) on its own (that is, 8-10% reduction in bit rate) or combined with spatial video image foveation (giving a 40-49% reduction in bit rate) was not found to be significant compared to spatial video image foveation alone. This was not surprising given that more image detail is removed in spatial video image foveation than in the temporal model.

From this analysis, it was concluded that the spatial video image foveation technique, developed to remove spatial redundancy in the peripheral field of view of BSL users, was the most effective way of achieving image compression. However, this initial assessment of potential bandwidth savings is limited since it does not consider the perceptual gains achieved by image foveation techniques which give priority to important image regions at limited transmission bandwidths. In general, reductions in bit rate come at the expense of higher levels of image compression in video images and this was reflected in the reduction in overall image fidelity indicated by the PSNR values obtained in the results (Table 9.7). However, PSNR is an automatic ('objective') measure of quality across the video scene and does not reflect the perceived (subjective) quality of the video sequences for the task of BSL communication where the face of the signer in the video is given priority.

Further subjective evaluation of pre-processed (spatially and temporally foveated) BSL video sequences, at a range of fixed bit rates, was conducted (Chapter 10) to determine the real benefits of this approach in terms of perceived (subjective) quality gain.

9.5 Summary

The subjective quality methods (TSCR and Binary Acceptability Method), developed in this thesis, were successful in obtaining end-user feedback on the perception of quality of video for BSL communication. The methods were applied to test user satisfaction with the video image content-prioritisation techniques described in Chapter 8. Spatial foveation of BSL video images was found to be acceptable up to a minimum contrast threshold (CT_0) equal to 0.1. Temporal foveation of BSL video images, using the FWTF method, was acceptable to BSL users at Filter Strength (FS) equal to 6.

Analysis of the potential coding gain of the content-prioritisation methods applied separately and together demonstrated significant bandwidth savings of up to 49 percent compared to the unprocessed video in a standard CODEC. The analysis identified the need for further subjective testing of the content-prioritised video to compare the performance of the standard and test systems at fixed bit rates. The perceived quality of BSL video at a range of fixed bit rates was tested to determine if selective content prioritisation of BSL video meets the subjective quality requirements for BSL video communication at lower bit rates than standard video compression schemes (Chapter 10).

	Unprocessed		CTo = 0.1			FS = 6			Combined CTo & FS		
QP	BR	PSNR Y	BR	PSNR Y	% BR saving	BR	PSNR Y	% BR saving	BR	PSNR Y	% BR saving
24	636.99	41.49	332.62	30.04	48	571.14	36.18	10	324.56	31.04	49
27	433.87	39.76	226.68	29.85	48	391.39	35.51	10	223.03	30.85	49
30	282.31	37.82	150.71	29.54	47	254.81	34.57	10	145.43	30.54	48
33	180.76	35.91	101.02	29.12	44	163.88	33.48	9	100.98	30.11	44
36	111.53	33.87	67.50	28.57	39	102.40	32.14	8	67.03	29.56	40

Table 9.7 Bit Rate (BR) and PSNR of Unprocessed and Pre-Processed (Spatially Foveated (CT_0 = 0.1), FWTF (FS = 6) and Combined Filters) Sequences Encoded at QP 24 to 36 and the Percentage Bit Rate Saving of the Processed Sequences Compared with the Unprocessed Video

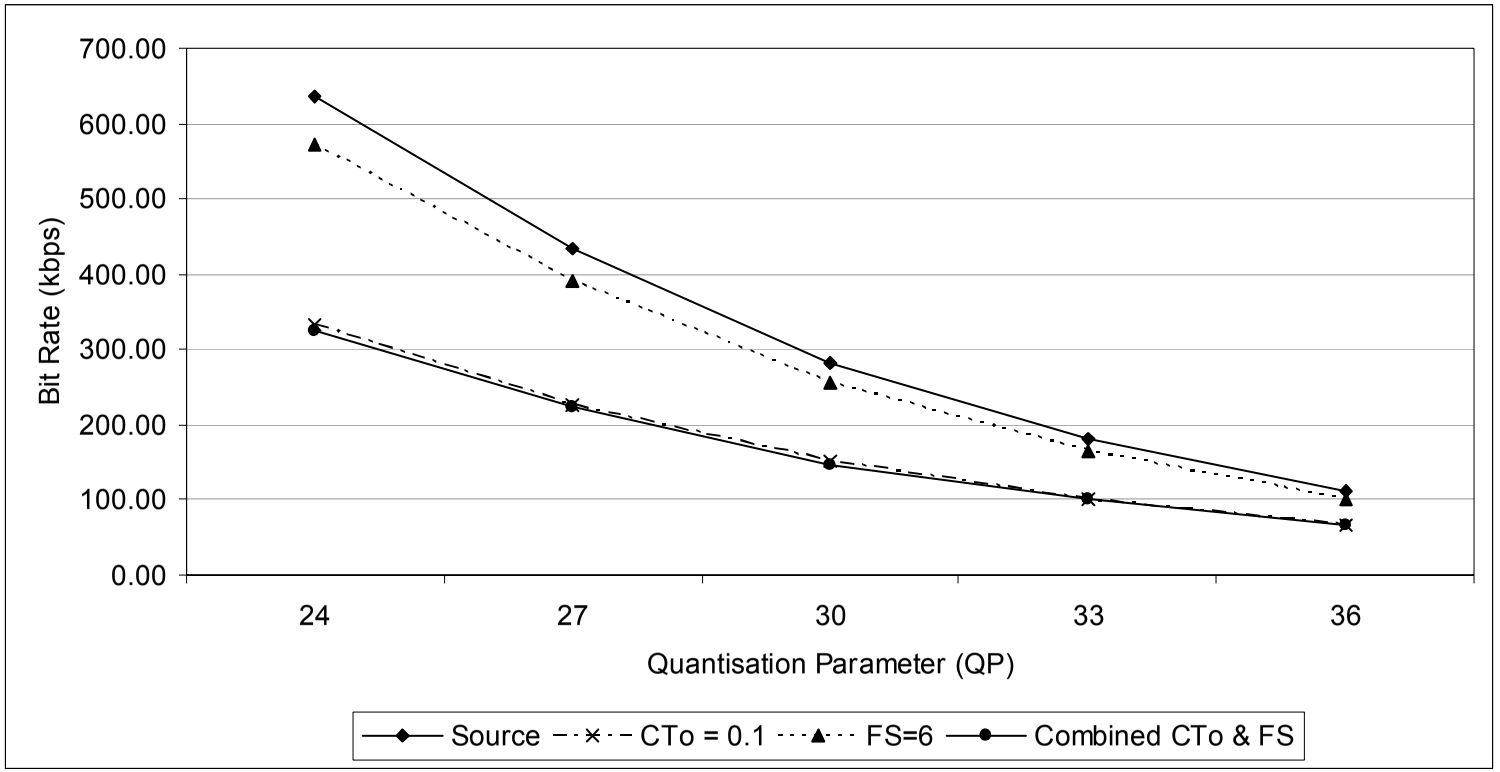


Figure 9.8 Comparison of Bit Rates for Pre-Processed and Source (Unprocessed) BSL Video at a Range of Quantisation Parameters (QP).

Chapter 10: Performance of Content-Prioritised BSL Video Coding

The perception of high quality video is generally dependent on the bit rate generated from the compression scheme; the higher the bit rate, the better the image quality (Sadka, 2002). In video communication systems the bit rate is restricted by the limited available transmission bandwidth. This requires a trade-off between bit rate and video quality resulting in a video communication system which does not meet the specific needs of the BSL user (Section 4.1). Optimised video compression (Section 4.2), based on the viewing behaviour of the user (Section 3.4) and the specific nature of the communication task (Section 3.1), offers the potential to reduce transmission bit rate requirements and achieve the perception of high quality. Eye tracking experiments (Chapter 6), video content analysis (Chapter 7) and subjective quality testing of visually optimised BSL video (Chapter 9) demonstrated that this could be achieved by selective prioritisation of BSL video image content by applying spatial video image foveation and FWTF techniques (Chapter 8).

Analysis of unprocessed and pre-processed BSL video material (Section 9.4) demonstrated that significant bit rate savings could potentially be achieved by the developed video image content-prioritisation schemes. The PSNR results obtained indicated that the outputs were of reduced quality in terms of image fidelity (Table 9.7). Objective measures have limited value in this application since they do not give an indication of user satisfaction for the task (as previously discussed in Chapter 5) and so an experiment was conducted to measure subjective quality as a valid indicator of video quality in this thesis.

The perceived quality of BSL video at a range of fixed bit rates was tested in an experiment designed to test the hypothesis that selective prioritisation of BSL video image content meets the subjective quality requirements for BSL video communication at lower bit rates than Standard video compression schemes. The Binary Acceptability Method (Section 9.1.2) for subjective quality evaluation was applied in this experiment to measure the gain in terms of perceived quality of pre-processed (content-prioritised) BSL video material compared to unprocessed video at a range of fixed bit rates. In addition, PC tests (Section 5.2.1.3) were applied, to identify user preferences in

situations where individual users gave equal/similar ratings to different test sequences and to provide the required level of test sensitivity (Table 5.1, Criteria 3) in the experimental design.

10.1 Experimental Design and Rationale

The aim of the experiment was to compare user satisfaction (acceptability) of BSL video sequences which had been pre-processed using spatial (video image foveation) and temporal (FWTF) filtering techniques with unprocessed material encoded at a range of fixed bit rates. The hypothesis was that the pre-processed BSL video material would be acceptable to users at a lower bit rate than the unprocessed material and that this perceptual gain[18] could be exploited in the design of video communication systems for deaf people using BSL. The experimental objectives were to:

1. Pre-process Test BSL video sequences by applying the spatial and temporal video image foveation techniques developed in this thesis (Chapter 8).
2. Encode the unprocessed and pre-processed test sequences at a range of fixed bit rates (from 50 kbps to 300 kbps).
3. Conduct subjective video quality testing to determine user satisfaction (acceptability) at each of the encoded bit rates.
4. Compare the subjective quality ratings for the unprocessed and pre-processed video material.

10.2 Method

A Binary Acceptability Method (Section 9.1.2) was applied to obtain subjective quality feedback from viewers of the original (unprocessed) and pre-processed unprocessed BSL video encoded at a range of fixed bit rates. This method was developed to measure quality in terms of acceptability for the user task (BSL communication) rather than a measure of overall degradation in image fidelity and enabled efficient capture of

[18] The perceptual gain is a measure of the improvement in the perceived quality of the test system compared to the standard system at low bit rates. It is measured in kilobits per second, this being the difference between the bit rate of the standard and test systems at the same (percentage) level of acceptability for BSL users.

concise user data in a controlled test environment for a range of test conditions. In addition, a Pair Comparison (Section 5.2.1.3) test was used to confirm user preferences.

10.2.1 Subjects

Subjective video quality assessment experiments were conducted with fourteen profoundly deaf-from-birth volunteers from the Aberdeen Deaf Social and Sports Club (ADSSC). Of the fourteen subjects six were male, eight were female and ages ranged from 37 to 80 years. For all subjects, British Sign Language (BSL) was their first language and English was a second language. For this reason all communications were in BSL, aided by a local British Sign Language Interpreter who was known to the participants.

Prior knowledge and experience of video telephony (for example, video telephones, video conferencing and Video Relay Services) has been demonstrated to have an impact on the expectations of quality of video communication systems (O'Malley *et al*, 1999) as previously discussed in Section 3.3. Three of the subjects participating in the experiment had prior knowledge/experience of video communication systems; one subject was a regular user of video telephony and two others had not used videophones but had seen product demonstrations. The remaining eleven subjects had no prior knowledge or experience of video communication systems.

10.2.2 Materials and Apparatus

The BSL video material for the experiment was selected from the original (unprocessed) test BSL Video Sequences created for the subjective quality tests in Chapter 9 (Table 9.2). Two short video sequences were selected to ensure that the test material contained a range of different BSL movements, expressions and gestures (Table 10.1).

Video Sequence	Number of Frames	Duration (Seconds)	English Translation of BSL Content
 Sequence 1 (Lisa School)	202	8.03	"I went to Aberdeen School for the Deaf. As I was growing up I used mostly signs and learned some oral communication."
 Sequence 2 (Lisa Television)	261	9.01	"When I watch TV, I look at the subtitles or the signer at the bottom right and I look back and to the picture."

Table 10.1 BSL Video Material (Unprocessed)

The test sequences are of sufficient duration to convey useful meaning in BSL while being short enough to satisfy experimental requirements to keep the overall test duration as short as possible for individual subjects. Short test sequences also minimise potential problems associated with short-term memory influences reported by Aldridge, Davidoff, Ghanbari, Hands and Pearson (1995), discussed in Section 5.2.2 and identified as a criterion for selection of test material (Table 9.1).

The first test sequence (Sequence 1) was processed using the spatial video image foveation method. This was achieved by stepping through, frame-by-frame, marking the central point for foveation (in this case the tip of the nose of the signer) and degrading the spatial quality from that central point according to the minimum Contrast Threshold (CT_0) and viewing distance. The CT_0 value for Sequence 1 was set to 0.1 (medium) as this was found to be the highest degree of spatial foveation which obtained a high Mean Opinion Score (MOS) in TSCR tests for the task of BSL communication (Section 9.1). The other parameters set in the foveation algorithm remained constant (viewing distance = 0.305, spatial frequency decay constant = 0.106 and half-resolution retinal eccentricity = 2.3). The output of the application of the spatial

video image foveation technique was video material with a smooth reduction in spatial resolution from the point of foveation (Figure 10.1).

(a) (b)

Figure 10.1 Frame 41 from Sequence 1(a) Unprocessed; (b) Foveated (CT_0 = 0.1)

The second video sequence (Sequence 2) was pre-processed to vary the temporal resolution in the video images using the FWTF method. The initial FWTF experiment (Chapter 9) determined the optimum Filter Strength value (FS = 6) up to which BSL video quality was considered to be acceptable. This value was set in the application of the FWTF method used to create the test material from Sequence 2 for encoding in this experiment. The output of the application of the FWTF technique was video material with reduced temporal resolution from the point of foveation (Figure 10.2).

(a) (b)

Figure 10.2 Frame 20 from Sequence 2 (a) Unprocessed; (b) FWTF (FS = 6)

The H.264/JVC reference software CODEC (JM.10.2) was used to encode and decode the unprocessed and pre-processed (foveated at CT_0 = 0.1 and FWTF at FS = 6) video

sequences at a range of fixed bit rates (50, 75, 100, 150, 200, 250 and 300 kbps) set in the encoder configuration file. The initial QP was set to the average QP for encoding the entire sequence to prevent deterioration of quality towards the end of the encoded sequences.

10.2.3 Procedure

The test BSL video sequences created for this experiment were displayed to the viewers on a seventeen-inch monitor with true colour, 32 bit display connected to a Dell Pentium IV PC with PCI Video Capture Card installed. Experiments were conducted under controlled conditions in a room with 100% artificial, overhead lighting. The subject was positioned at a comfortable viewing distance (four to six times the screen height) from the PC monitor and was given instruction in BSL through a qualified BSL Interpreter. All communication with the subject was in BSL, the subject's first language. No printed instructions or feedback forms in English language were used. A short training session (containing seven video sequences) was conducted with each individual subject prior to the main experiment to familiarise the participants with the experimental procedure.

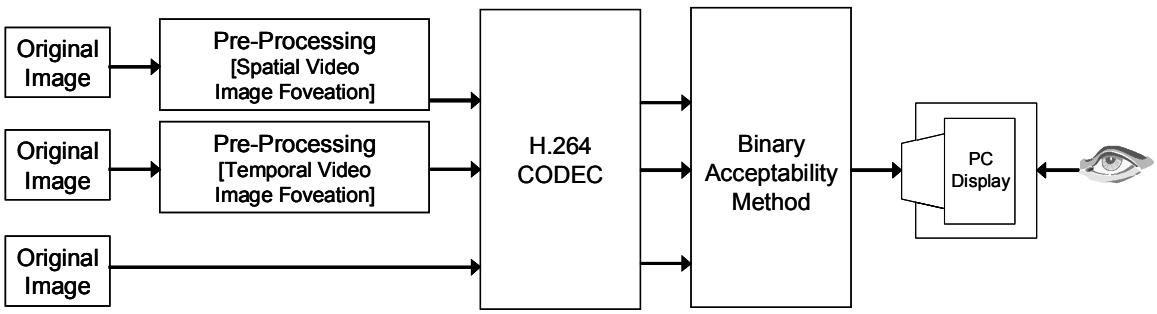


Figure 10.3 Block Diagram of the Comparison of the Performance of the Standard and Content-Prioritised (Spatial and Temporal Image Foveation) Systems at Fixed Bit Rates

Each subject was asked to view a series of a total of 28 single BSL video sequences, which had been encoded at a range of bit rates (50, 75, 100, 150, 200, 250 and 300 kbps). The test material was presented in four groups (Tests A, B, C and D) in order of

increasing bit rate (Table 10.2a). The video sequences were displayed full-screen, one-at-a-time and each subject was asked to report if the video was acceptable or not acceptable for BSL communication during display of a plain grey-coloured display image at the end of each sequence. Subjects had the option to request that a sequence was repeated if required.

An additional Pair Comparison (PC) test was conducted, at the end of the Binary Acceptability Method test, to confirm the preference of each subject in a comparison between the same unprocessed and pre-processed sequences at a fixed bit rate of 150 and 200 kbps. In this test, the video sequences in each pair were displayed to the user as full-screen images, separated by a plain grey-coloured display image which was displayed for a two second period. Each pair of video sequences was separated by a plain grey-coloured display image which was displayed for a ten-second period. The subject expressed preference for either the first or second sequence in the pair during this period. The order of the video sequences in the pair was varied during the experiment to ensure that the correct number of repetitions were included for this test (Table 10.2b).

Test A: Sequence 1 (unprocessed)	
Video Number	Bit Rate (kbps)
1	50
2	75
3	100
4	150
5	200
6	250
7	300

Test B: Sequence 1 (CTo = 0.1)	
Video Number	Bit Rate (kbps)
8	50
9	75
10	100
11	150
12	200
13	250
14	300

Test C: Sequence 2 (unprocessed)	
Video Number	Bit Rate (kbps)
15	50
16	75
17	100
18	150
19	200
20	250
21	300

Test D: Sequence 2 (FS = 6)	
Video Number	Bit Rate (kbps)
22	50
23	75
24	100
25	150
26	200
27	250
28	300

Table 10.2a Test BSL Video Sequences (Twenty-Eight BSL Video Sequences Created from the Unprocessed and Pre-processed Material at Bit Rates from 50 to 300 kbps) for the Binary Acceptability Test

PC Test	Test Video Sequence Pairs	
	Video Number	Video Number
1	19	26
2	18	25
3	11	4
4	4	11
5	26	19
6	12	5
7	25	18
8	5	12

Table 10.2b Test BSL Video Sequences for the Pair Comparison Test.

10.3 Results

For Sequence 1, binary acceptability data were obtained from deaf participants for unprocessed and pre-processed (foveated) video material at fixed bit rates (Table 10.3a). Acceptability levels were found to be significantly higher (7-57% greater) for the pre-processed (foveated) video material than for the unprocessed video at low bit rates (below 200 kbps), shown in Figure 10.4a[19]. One hundred percent acceptance is achieved at 200 kbps for the pre-processed video and at 250 kbps for the unprocessed material demonstrating a perceptual gain of 50 kbps. These results were confirmed in the Pair Comparison tests (Table 10.3c); all subjects preferred the spatially foveated sequence (Video Number 12) to the unprocessed material (Video Number 5) at 200 kbps and all but one subject (92.86%) preferred the spatially foveated sequence (Video Number 11) to unprocessed material (Video Number 4) encoded at 150 kbps (this subject reported that he was not able to distinguish between the quality of these sequences in this test).

The results for Sequence 2 (Table 10.3b) also indicate significantly higher acceptability (7-28% greater) for pre-processed (FWTF) video at bit rates less than 200 kbps. Unanimous acceptance of the pre-processed video is only achieved at the same bit rate for the unprocessed material and so perceptual gain is not evident at this (100%) acceptability level for FWTF video (Figure 10.4b). The Pair Comparison (Table 10.3c) test confirmed preference (by 92.86% of subjects) for the FWTF sequence (Video Number 26) to the unprocessed (Video Number 19) coded at 200 kbps. Nine of the fourteen subjects (64.29%) preferred the FWTF video (Video Number 25) to the unprocessed material (Video Number 18) at 150 kbps.

None of the subjects reported problems in understanding the BSL in the test sequences. Ten of the subjects specifically commented that the clarity of the BSL video content was improved when the face of the signer was of acceptable quality. One of

[19] Error bars in Figure 10.4 are included at the 95% confidence interval.

the subjects (female, age 44 years), who was a regular user of video telephony (Leadtec IP Broadband Videophone) in her employment, rated the unprocessed video as unacceptable up to 300 kbps and found the pre-processed sequences to be acceptable at 150 kbps for foveated video and 200 kbps for FWTF video. This demonstrated a significantly higher perceptual gain (150 kbps for foveated video and 100 kbps for FWTF video) for this participant. The remaining subjects, who had little or no experience of video telephony, tended to accept the quality of the unprocessed video at lower bit rates and had personal perceptual gains of around 50 kbps which represented the findings for the sample group.

Bit Rate (kbps)	Unprocessed (% acceptable)	Foveated CTo = 0.1 (% acceptable)
50	7.14	14.29
75	14.29	35.71
100	14.29	71.43
150	42.86	92.86
200	92.86	100.00
250	100.00	100.00
300	100.00	100.00

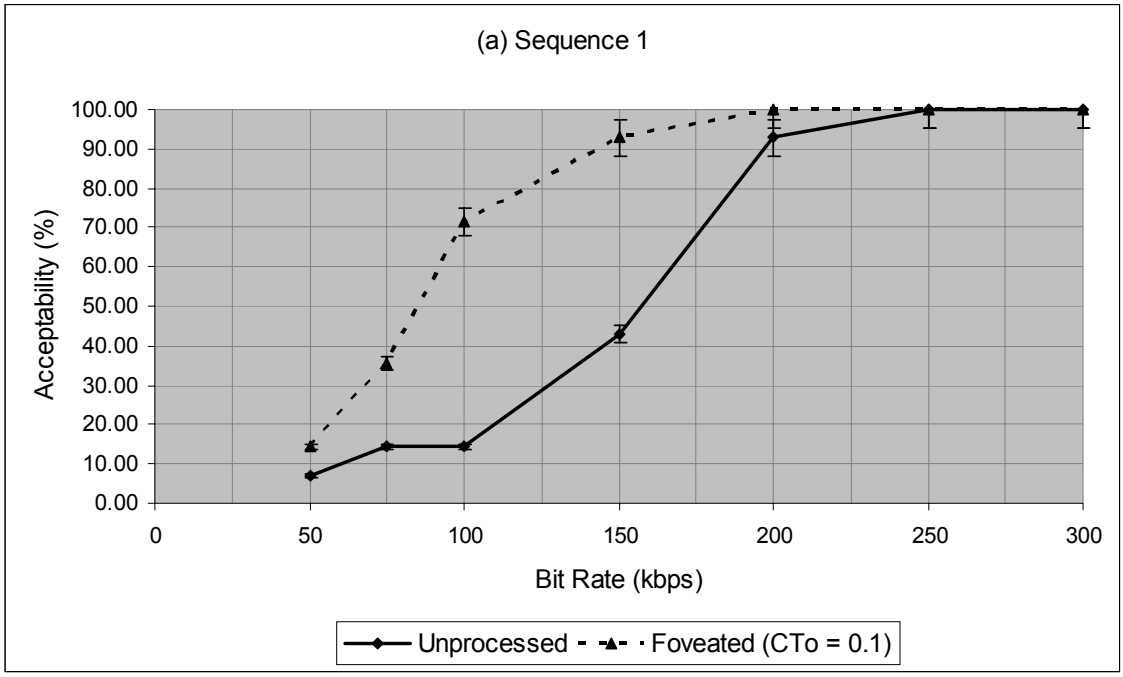
Table 10.3a Acceptability of Sequence 1(Unprocessed and $CT_0 = 0.1$) at Fixed Bit Rates

Bit Rate (kbps)	Unprocessed (% acceptable)	FWTF FS = 6 (% acceptable)
50	7.14	7.14
75	7.14	21.43
100	14.29	42.86
150	35.71	64.29
200	85.71	92.86
250	92.86	92.86
300	100.00	100.00

Table 10.3b Acceptability of Sequence 2 (Unprocessed and FS = 6) at Fixed Bit Rates

PC Test	Video 1		Video 2	
	Video Number	Preference %	Video Number	Preference %
1	19	7.14	26	92.86
2	18	35.71	25	64.29
3	11	92.86	4	7.14
4	4	7.14	11	92.86
5	26	92.86	19	7.14
6	12	100.00	5	0.00
7	25	64.29	18	35.71
8	5	0.00	12	100.00

Table 10.3c Pair Comparison of Unprocessed and Pre-processed Sequences



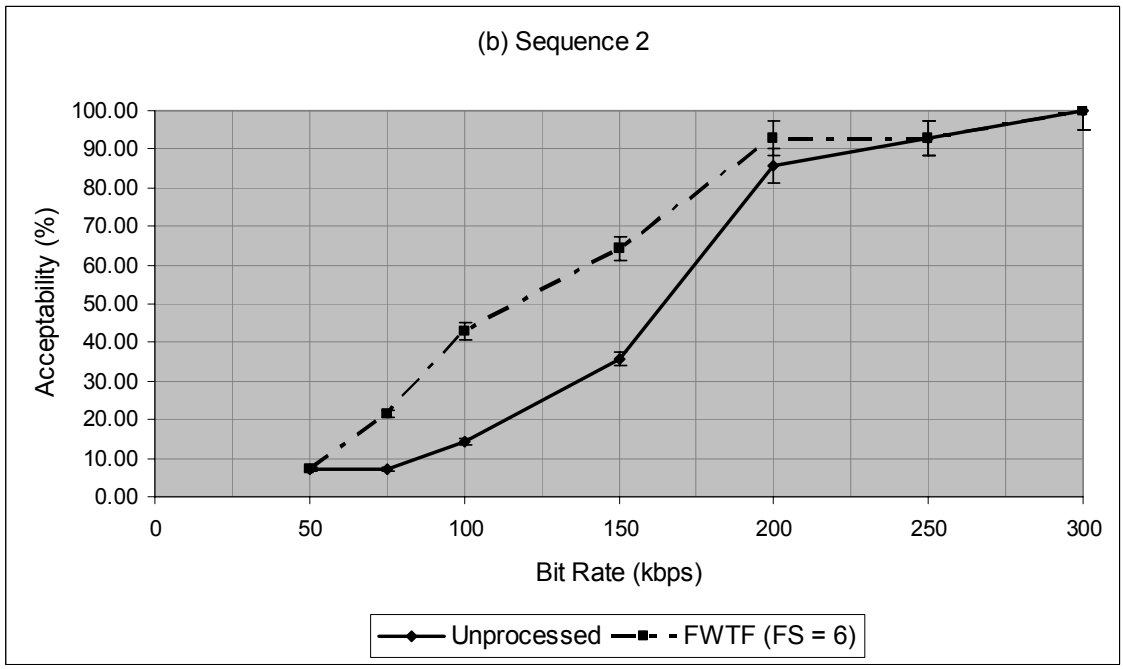


Figure 10.4 Acceptability Plots for Unprocessed and Pre-Processed Video for (a) Sequence 1 and (b) Sequence 2

10.4 Discussion

BSL video which has been pre-processed to prioritise important image content based on the visual response of deaf people was demonstrated to be more acceptable for BSL communication than the unprocessed material over a range of fixed bit rates. These results demonstrate significant perceptual gains (of 50 kbps) at one hundred percent acceptability levels for pre-processed (foveated) BSL video. However, it could be argued that a lower acceptability rate can be tolerated in the design of systems (McCarthy, Sasse and Miras, 2004). In this experiment, significant gains for the pre-processed video were observed at 70% acceptability (75 kbps) for foveated video sequences and at 50% acceptability (50 kbps) for FWTF video sequences. The Pair Comparison tests confirmed preference for the pre-processed video material compared to the unprocessed video at 150 and 200 kbps. In general, the results demonstrate significant perceptual advantages of spatial and temporal pre-processing based on HVS models (particularly spatial video image foveation) for this user group. The perceptual gain was greatest at bit rates between 50 kbps and 200 kbps and was highest for the experienced video telephony user in the sample group. The low expectation of quality (indicated by accepted quality of low bit rate encoded unprocessed video) from subjects with little or no experience of video telephony meant that the overall gain afforded by the pre-processed material was reduced for the subject group as a whole.

BSL video material which had been pre-processed based on visual behaviour mechanisms was found to be acceptable to users at lower bit rates than the unprocessed material. This perceptual gain can be exploited in the design of video coding algorithms in communication systems for deaf people using BSL communication. Selective prioritisation of important regions of the video image enabled more efficient video compression and improved the perceived quality of the video content for BSL communication.

PART THREE: DISCUSSION AND CONCLUSIONS

Chapter 11: Discussion

This chapter reviews the research problem addressed in this thesis and discusses the development of content-prioritised video for sign language communication and the potential for further work.

11.1 Review of the Research Problem

British Sign Language (BSL) is a highly structured visual linguistic system which expresses vocabulary and grammar spatially and temporally using multi-channel signs and gestures, including facial expression (such as eyebrow movement, eye blinks, eye gaze location and mouth/lip shapes), head, hand and body movements and finger-spelling (Section 3.1). The communication of information between deaf people using freely expressed BSL is rapid, complex and places significant demands on visual media applications such as video telephony and video conferencing (Section 3.2) which are regarded as enabling technology for deaf people using sign language. Video compression is required to reduce the amount of digital video data (bits) for transmission and storage in a communication system by removing subjectively redundant information (Section 4.1). Although there have been significant increases in available bandwidth for networked communications in recent years, there remains a requirement for efficient use of Internet and broadcasting bandwidth resources. This is the case for video communication where the perceived quality of service will be of particular importance on high tariff, high bandwidth services. It is especially true for the provision of real-time video communication of BSL delivered over networks and for the future development of 'closed signing' (discussed in Section 1.5) over broadcasting channels. At low communication bit rates (that is, less than 200 kbps), video compression algorithms optimised to meet the needs of the user can be developed to improve perceived video quality. Standard systems with bit rates of 256 kbps currently giving 'good quality' quarter-screen (CIF) video could also be optimised to provide good full-screen DVD quality video images.

Generally, video communication systems are designed based on the sensitivities and limitations of the HVS (Human Visual System) and assumptions about viewing behaviour. In this thesis, the creation of a system designed for the specific needs of

deaf people, using video media applications for personal communication, was based on knowledge and understanding of *where* deaf people look and *what* deaf people see during a BSL video communication task. These human factors were determined by experiment and analysis and informed the development of novel methods to optimise and measure quality.

11.2 Review of Experimental Work and Developments

This section reviews the research and development of methods to measure BSL video quality, investigate the visual responses of BSL viewers, analyse BSL video image content, design a video coding scheme based on the visual response to BSL video content and measure performance of the content-prioritised video (compared to standard systems) for BSL communication.

11.2.1 Video Quality Assessment

Video quality assessment of BSL video communication is of critical importance in the design of a system which aims to meet the specific needs of deaf people. The development and application of an appropriate method of assessing video quality was a significant challenge in this thesis. Automatic 'objective' evaluation techniques (which apply mathematical models designed to measure the human visual response in terms of criteria and metrics that can be measured objectively) were found to be inappropriate since they measure image fidelity rather than the perception of quality for a visual task (Section 5.1).

The user-centred design approach in this thesis required a method of video quality assessment which accounted for the human cognitive and perceptual factors affecting acceptability for the task (BSL communication). Subjective information from users was obtained using techniques developed to assess user satisfaction for task performance. The development of these techniques was based on general design principles (Table 5.1) and specific design criteria (Table 9.1) identified for the experimental work in this thesis.

Standard methods of rating subjective video quality assessment for multimedia applications (ITU-T Recommendation P.910, 1999) and for television pictures (ITU-T

Recommendation BT.500-11, 2002) were found to be unsuitable for assessing quality for the BSL communication task (Section 5.2.2) since these methods were designed for measuring image fidelity, using rating scale descriptors for the judgement of impairments. Particular problems were identified with the rating scale descriptors; the difficultly of translating the rating descriptors into different languages (including BSL) and the difficulty for users in making decisions based on the standard rating scales. New methods (based on the standard ITU recommendations) and alternative (non-standard) approaches were developed and applied in this thesis.

In the subjective quality assessment of spatially foveated BSL video (Section 9.2) a Task-Specific Category Rating (TSCR) scale (Table 9.3) was developed for a double stimulus method of subjective quality assessment based on the Degradation Category Rating (DCR) method of ITU-T Recommendation P.910 (1999).

The rigour of the testing regime for the standard DCR method and adaptation of the original five-point rating scale to specify the criterion for assessment at each point on the scale for the BSL communication task, provided relevant detailed feedback which was statistically tested for consistency of response in a comparison of the unprocessed and foveated video material. However, although the assessment criteria and experimental procedure were successful in obtaining valuable information, a number of limitations of this method were identified which were addressed in subsequent subjective assessments with deaf participants.

Although the deaf participants did not complain of experimental fatigue, it was felt that the number of repetitions required for ensuring consistency of user response in the testing regime made the application of this approach tedious and time-consuming for the subjects and the BSL Interpreter. It was also felt that the requirement to make a decision about quality on a five-point scale of detailed descriptors in direct comparisons with the reference material (in the double stimulus approach) contributed to the time and effort required from participants. A direct comparison was useful for assessing performance of the test material with reference to the original unprocessed material in terms of perceived quality for the task. However, a single stimulus approach (including presentation of the unprocessed material) would have provided sufficient evaluation of user satisfaction, enabled comparison of responses to unprocessed and pre-processed video material at the analysis stage and been more comfortable for the participants.

A Binary Acceptability Method (Section 9.1.2), based on the assessment approach by McCarthy, Sasse and Miras (2004) and on recommendations in the ITU standards (for the duration and presentation of the video material), was developed to address these limitations, while ensuring that the experimental design criteria were met. This method was applied in the subjective quality assessment of temporally foveated BSL video (Section 9.3) and in the final evaluation of the encoded content-prioritised BSL video (Chapter 10). The Binary Acceptability Method was successful in reducing the time and effort required from the participants by presenting the video material as a single stimulus in order of decreasing/increasing strength of the test condition (to allow ease of decision making about acceptability) and by reducing the assessment criteria to a binary decision (acceptable and not acceptable for BSL communication). In addition, the Pair Comparison method (5.2.1.3) from ITU-T Recommendation P.910 (1999) was applied to a selection of the test material in the final evaluation (Chapter 10) to confirm user acceptability preferences and provide additional assurance regarding the results at 150 and 200 kbps.

The process involved in the development and implementation of a quality assessment method which met clearly defined experimental criteria was a significant contribution to the research. It led to confidence in the results obtained from experimentation which relied on the subjective opinions of a relatively small but acceptable number of users. It was concluded that assessment of the perception of video quality for a specific communication task requires clear user assessment criteria (using a simple binary decision or a more detailed multi-point scale of task-specific rating descriptors) and an user-centred approach to experimental design.

11.2.2 Eye Movement Tracking

Eye Movement Tracking (EMT) methods were developed to investigate the visual responses of BSL viewers. EMT experiments determined *where* deaf people look during BSL communication and which region of the BSL video content is most important to the viewer. The link between eye gaze and visual attention is assumed to be strong due to the complex nature of the visual task (Section 2.2).

Specific technical limitations and methodological issues (Section 6.2) associated with the use of EMT for obtaining physical measurements of eye gaze from deaf subjects

during a BSL video communication task were addressed in the design and implementation of this technique. Problems with calibration of individual subject's eye position before eye tracking, reported by other researchers (Duchowski, 2003; Hyőnä and Lorch, 2004; Jacob, 1991 and Schnipke and Todd, 2000) using this method, were encountered in this thesis (Section 6.4.2.1). It was necessary to discard data from subjects whose eye movements could not be accurately calibrated and tracked to ensure reliability and robustness of the results. However, in order to maximise the value of the information obtained from successful eye tracking of subjects, problems associated with the complexity of data collection and analysis of eye tracking data, reported by Duchowski (2003), Salvucci and Goldberg (2000) and Rayner (1998), were addressed. This was successfully achieved by developing algorithms (in Matlab) for visualising EMT data superimposed on the BSL video image stimulus, establishing procedural rules for identifying fixations on designated regions of interest and by detailed frame-by-frame analysis of eye gaze coordinates with respect to the BSL content in each frame, resulting in the production of eye movement timelines for each subject (Section 6.4.3.3).

The EMT analysis established a consistent characteristic pattern of viewing behaviour of deaf people engaged in the active task of BSL video communication. The results demonstrated that the deaf viewers foveated mainly on the face of the signer and did not track the movement of the hands and fingers during a BSL communication. It was concluded that the face of the signer in the video image was the most important region for BSL communication and that the hands and finger gestures, which are also important parts of the sign, are viewed in lower resolution peripheral vision when they are not close enough to the face to be seen in high resolution vision afforded by the fovea of the eye around the point of fixation (Section 6.4.4). By viewing the face region of the signer in foveal (high resolution vision) the HVS gives priority to that region of the image for sign language communication.

11.2.3 BSL Video Analysis

Analysis of the content of the BSL video material revealed the characteristics of the BSL video material in terms of the optical flows in different regions of the BSL video

images and also in terms of the priority which these regions were given in a standard CODEC (Chapter 7).

Optical flow analysis of the BSL video material (used in the EMT and subjective quality testing experiments) enabled the identification of *what* deaf people were seeing in foveal and peripheral vision in terms of image motion flows (Section 7.1). The face of the signer in the BSL video images, observed by deaf people in the foveal (high resolution) field of vision, was characterised by small motion flows. These fine optical flows are associated with small, rapid changes in contrast (high frequency image components) in the video sequence. Significant motion flows (corresponding with low spatio-temporal frequencies or changes in luminance over a large area) occurred in the region of the BSL video image which included the movement of the arms and hands. This region was not tracked by the eye in the EMT experiments and was viewed in low resolution (peripheral vision) during BSL video communication.

Analysis of the bit allocation of a basic H.263 encoder revealed the priority given to each region of the BSL video images in the standard coding scheme (Section 7.2) in terms of the proportion of available coded bits allocated to each region. Coding priority (43-47% of the total bit count) was given to the arms/hands region which contained the greatest motion flows in the BSL video sequences. The most important region of the image for BSL communication (the face of the signer) was given low priority in the coding scheme (8-11% of the total bit count) which was not significantly greater than the allocation given to the image background (3-7% of the total bit count) and less than the 14% bit allocation to the stationary background object in test BSL Sequence 1. The arms/hands region was allocated most bits in the encoder, yet this region was viewed only in peripheral (low resolution) vision suggesting that high resolution detail is not required in this region. This theory is supported by studies of object perception performance (Section 2.1.3) by van Diepen, Wampers and d'Ydewalle (1998). Van Diepen *et al* (1998) found that the perception of objects in peripheral vision was improved by high pass filtering (to remove the low spatial frequencies).

In summary, the priority given by the video encoder (to the large movements of the hands) did not match the priority given by the human eye to important regions of the video image for sign language communication (that is, the small detailed movements in the face). This demonstrated the potential for improving the perception of quality and

reducing bandwidth requirements of BSL video by giving priority to the most important region of the video image for BSL communication in the video coding scheme.

11.2.4 Content-Prioritised Video Coding for BSL communication

Based on the analysis of visual behaviour and video content, it was concluded that a suitable strategy for coding BSL video would be to prioritise video image regions selectively in order to provide high quality in the face region and reduced resolution of the arms/hands of the signer in the video. Options for content prioritisation included object-based coding, macroblock-level prioritisation and pre-processing. Object-based coding (for example, using tools provided by MPEG-4 Visual Main and Core Profiles) arguably gives the highest flexibility in coding foreground and background objects with different coding parameters and refresh rates. However, accurate segmentation along object boundaries is computationally intensive and there is a lack of practical implementations of the object-based tools within MPEG-4 Visual. A more practical approach is to prioritise macroblocks selectively during coding with a block-based scheme such as MPEG-4 Visual Simple Profile or H.264. One approach is by controlling the quantisation parameter (Section 4.1.3.4) to provide varying spatial quality in different regions of the scene. However, this segmentation approach can only support prioritisation of approximate regions.

Pre-processing (Section 4.1.3.1) of the video image to filter spatial detail prior to encoding is an alternative approach to selective coding. Foveated video image processing was the approach developed in this thesis to give priority to the face of the signer in BSL video communication at the pre-processing stage of the video coding process. This method was developed to exploit the visual response of deaf people to BSL video and the foveal response of the HVS.

The foveal response of the HVS is the human mechanism by which ‘important’ content is given priority in a visual scene. Information captured by the fovea is viewed in high resolution and this declines with foveal eccentricity in the peripheral field of view (Section 2.1.1). Image foveation techniques mimic the human foveal response and produce image compression through reduction of spatial information. The image foveation algorithm developed by Geisler and Perry (1998) modelled the decline in spatial resolution away from the point of gaze using a multi-resolution pyramid (Section

2.3.2) and psychophysically measured parameters (Table 8.1). This technique made optimal use of foveation by removing spatial detail which could not be resolved by the HVS. However, Geisler and Perry (1998) determined that “in practise foveation only adds a little to the total compression because the motion is primarily confined to the foveated (i.e. ‘clear’ or non-blurred) region”. They concluded that the value of image foveation was in video where there are small regions which require detailed inspection or where “motion is over extended image regions” and there are severe bandwidth limitations. They assumed that the viewer foveates on the image region which contains most motion but this was not found to be the case for deaf viewers in the EMT and video content analysis of BSL video in this thesis. Geisler and Perry (1998) also identified that practical application of this method was limited by the requirement for knowledge of the point of foveation. Such schemes generally require EMT but in this thesis it was argued that in applications where gaze location can be predicted for specific tasks, such as BSL communication, models of visual behaviour are appropriate. It was postulated that, with knowledge about the viewing behaviour of deaf people obtained from EMT and by raising the degree of foveation away from the face region of the signer in BSL video (by increasing the minimum contrast threshold in the image foveation algorithm above psychophysical limits), a higher degree of image compression would be achieved without affecting the perception of quality. A method of image foveation was developed for this thesis which implemented the algorithm of Geisler and Perry (1998) for processing BSL video sequences at variable minimum contrast threshold values (to set the degree of foveation ‘blurring’) and viewing distances to control the size of the ‘clear’ foveated region (the face of the signer in the BSL video). This method was successful in producing pre-processed BSL video which displayed smooth degradation in spatial resolution increasing with distance away from the face of the signer at a range of minimum contrast threshold values (Chapter 8).

An additional method was developed to apply the foveation model to vary temporal resolution in BSL video images. This was based on the hypothesis that, as with spatial resolution (that is, given the established viewing behaviour of deaf people engaged in a BSL communication task), uniform temporal resolution across the whole video scene would not be required for BSL communication. A novel method of temporal resolution filtering, Foveation-Weighted Temporal Filtering (FWTF), was developed and applied to reduce the temporal resolution of BSL video away from the region of importance (the

face of the signer in the video) at a range of filter strengths (Chapter 9). This method was successful in achieving a foveation model of temporal resolution in BSL video sequences.

11.2.5 Performance

Subjective quality evaluation of spatially and temporally foveated BSL video sequences, processed at a range of filter strengths, established the limits for user acceptability in BSL video communication (Chapter 9). A minimum contrast threshold of 0.1 for spatially foveated video (Section 8.3) and a filter strength value of 6 for temporally foveated material (Section 9.3) produced the maximum degree of foveation (blurring) in the peripheral field of view which provided acceptable video quality for communication of BSL.

Objective analysis of the encoded pre-processed BSL video compared to the unprocessed material (Section 9.4) determined potential savings in bandwidth requirements of the foveated video image sequences. The most significant gains were obtained from the application of spatial video image foveation (that is, a reduction in encoded bit rate of 39-48% compared to the unprocessed video).

The perceptual gain of video image foveation, measured in terms of user acceptability for the task, was determined by subjective quality evaluation of the encoded unprocessed and pre-processed (foveated at the maximum filter strengths determined for user acceptability) video material at a range of bit rates set in the encoder (Chapter 10). Foveated BSL video sequences were found to be acceptable at lower bit rates than the original unprocessed material. This was particularly evident at bit rates between 50 and 200 kbps. Perceptual gains of 50 kbps at 100% acceptability levels and 75 kbps at 75% acceptability levels were obtained.

It was observed that the experience of the deaf viewers had an impact on the results since users with no experience of video telephony accepted the quality of the original video material at low bit rates and this limited the gain afforded by the pre-processed (foveated) video material.

11.2.6 Summary of Developments

The image foveation techniques developed in this research, applied at the pre-processing stage of video coding, were successful in achieving the perception of quality for BSL communication at low bit rates by giving priority to the most important video image content for the task. The methods developed in this thesis contributed to the successful implementation of a real-time application of perceptually optimised video for sign language communication (and for video conferencing and surveillance applications where the region of importance is known and can be tracked) in a Proof of Concept project funded by Scottish Enterprise.

11.3 Further Work

The foveation algorithms were not applied in real-time in this thesis as the purpose of their development was to identify their potential for increasing coding efficiency and the perceived quality of the outputs. A real-time implementation of the spatial video image foveation algorithm was developed in a Proof of Concept project arising from this research. The real-time implementation uses face detection and tracking algorithms to automatically detect the central point of foveation. The computational complexity of the algorithm was reduced in the real-time implementation to improve the performance of the developed system for commercial application.

In addition to the development of a real-time application in the Proof of Concept project a number of opportunities were identified to improve video communication of visual information for BSL users in this thesis.

Mutual gaze was identified as an important factor in real-time video communication applications including BSL video communication and general video conferencing applications (Section 3.4). Mutual gaze is an objective indicator of communication behaviour (O'Malley *et al*, 1999) and eye contact is known to be important for effective sign language communication. Conventional video conferencing systems do not support natural mutual gaze due to vertical disparity between the camera and the image of the receiver's eyes and so gaze awareness may be a factor for investigation in the future development of real-time video communication systems for deaf users and in wider video conferencing applications.

Another opportunity for further research is in the application of content-prioritised coding of BSL to the development of closed signing in digital television broadcasting (discussed in Section 1.5). Bandwidth is a scarce resource in broadcasting communication channels and so there is an opportunity to develop highly efficient coding schemes based on visual behaviour to signed content on television to significantly reduce the number of bits required to represent the visual information. Study of the visual responses to the signed component and the general picture in the broadcast would enable this service to be optimised to provide high quality within limited bandwidth resources.

In addition to further work in different application areas, further research is proposed to identify a new approach to giving priority to visually important video content. The general approach taken to prioritise important image content in video image sequences in this thesis was based on ROI coding. This approach requires detection and tracking of the regions of interest or importance in the video coding scheme. An alternative approach was identified from the EMT and video analysis conducted in this work which places the emphasis on the type of image content rather than the location of the content in the video sequence. This new approach is introduced in this section of the thesis as an alternative model for content-prioritised BSL video communication and as a suggestion for further work.

The EMT experiments identified the gaze locations of deaf people viewing BSL video (principally the face of the signer in the video). Knowledge of BSL and the observations of Siple (1978) suggest that experienced BSL users look at the face to pick up fine detailed motions associated with facial expressions which are crucial to the comprehension of the language. Analysis of the motion flows in the BSL video sequences, applied in the experimental work for this thesis, revealed that the foveated region contained small optical flows. Analysis of the bit allocation of this region identified that these high frequency components were not given priority in a standard-based video CODEC which compresses images by discarding some of the high frequencies which are assumed to be ‘less important’. This suggested that the design of a video communication system for deaf people using BSL, which gave priority to the face of the signer, would need to preserve some of the higher frequency components of the image which are typically removed in a standard video CODEC design at the

expense of low frequency components not tracked during BSL communication. This could be achieved in the DCT (Discrete Cosine Transform) of a block-based CODEC. The DCT (Section 4.1.3.3) is a tool for removing subjective redundancy by prioritising higher order bits so that the quantiser removes lower order bits prior to encoding. Previous researchers have applied DCT-based compression techniques in 'region of interest' coding schemes. Sheikh, Evans and Bovik (2003) used DCT sub-band weighting to suppress high frequency components (assumed to be visually less important for general viewing) in segmented image regions determined by an 'approximate' foveation model. However, this approach produced blocking artefacts and less suppression of high frequency components than an image foveation filter. Porikli (2004) used texture characteristics from DCT coefficients and available motion information to generate an object partition tree for object segmentation (for compression, indexing search and content retrieval) of encoded video images to reduce computational load. Porikli (2004) based his approach on the fact that DCT coefficients have a relationship with spatial frequencies and that different components of the image have different subjective importance.

This thesis determined that deaf viewers do not adopt the general model of viewing behaviour applied in standard video CODECs. It is postulated that a video CODEC designed to retain visually important high frequency image components at the expense of non-important low frequency components would offer the perception of high quality for BSL communication and potentially reduce the bandwidth requirements. This could be achieved by selective weighting of the DCT coefficients to give priority to important frequencies (to be determined by experimentation) for subsequent quantisation and encoding. Prioritisation of important video image frequency components for the task is an alternative approach to content prioritisation which requires further work to establish the frequency thresholds (and how they could be separated from background frequencies), weightings and impact on perception of quality for the deaf viewer.

Chapter 12: Conclusions

The research presented in this thesis was successful in developing methods for content-prioritised sign language video communication which achieved improved perception of quality at low bit rates (below 200 kbps) in subjective testing by deaf viewers. A subjective quality assessment method was created which addressed design principles and criteria identified for the experimental work and for deaf participants. This chapter reviews the achievements of this thesis in terms of the research objectives and contributions to the body of knowledge.

12.1 Review of Research Objectives

This section evaluates the theoretical and methodological developments and research contributions which were achieved for each of the research objectives described in Section 1.3.

12.1.1 Human Factors and Design Constraints Affecting the Development of Video Communication Systems for Deaf People using BSL

Evaluation of the human perceptual and cognitive factors affecting the communication of BSL (Chapter 3) established that knowledge and understanding of the visual response to BSL was critical to the design of video communication systems which meet the needs of deaf people. According to Siple (1978), sign language has developed to maximise communication of visual information within the limitations of the HVS. This led to the conclusion that a video communication system designed to meet the visual requirements of deaf people must also exploit the specific limitations and processing capabilities of the HVS for the BSL task. A number of common misconceptions and assumptions about sign language were identified which had methodological implications for this work (Section 3.1). Sign languages, including BSL, are complex linguistic visual systems, with their own grammar and syntax which have national and regional variations similar to spoken languages and dialects. This meant that standard test sequences were not appropriate for testing responses to BSL video communication and so original BSL video material (in local BSL) was created to ensure

that there were no barriers to the BSL communication task apart from the perceived quality of the video images. Assumptions about the relative importance of the face and hands regions for BSL communication in previous research led to the development of systems which did not consider the real priorities given to BSL video content by the HVS. This thesis established the mechanisms of the HVS which give priority to important content in the visual scene (Chapter 2). Knowledge of the foveal response of the HVS, which gives priority (in terms of image resolution) to visually selected image content around the point of gaze, contributed to the design of a system which aimed to prioritise visually important information. In addition, understanding the response of the HVS to spatial and temporal information, the mechanisms for processing image features and objects and the superior image processing capabilities of experienced sign language users in terms of face discrimination and peripheral vision was critical. The consideration of these human factors led to the informed design of methods to prioritise visually important content for BSL communication, provide perceptual gains for the user and reduce the amount of information which required to be transmitted in a video communication system.

Investigation of the design factors influencing the development of a video communication system for deaf people identified the limitations of current systems and the options for optimising video compression schemes (Chapter 4). Evaluation of human and technical factors affecting the design of video communication systems for deaf people contributed to the development of methods of content prioritisation for BSL video communication optimised to meet the needs and capabilities of users within design constraints.

12.1.2 Visual Response of Deaf Viewers to BSL Video Content

Investigation of the visual response to BSL video content by profoundly deaf subjects was conducted in Eye Movement Tracking (EMT) experiments (Chapter 6). Methodological problems, highlighted by previous researchers in terms of the accurate capture and analysis of eye gaze data, were addressed. Although similar problems were encountered with accurate calibration (and therefore subsequent capture) of eye movements which reduced the data set available, contributions were made in the analysis of data which ensured that the maximum value was obtained from the

available data. The EMT data were visualised with respect to the video image content on a frame-by frame basis by superimposing the eye gaze coordinates on the video frames. The results were analysed with respect to designated regions of the video image by creating a timeline, for each subject, which plotted the gaze region against the image regions containing active BSL content (for example, movement of the hands and facial expressions) during the BSL sequence. The development of these techniques enabled detailed analysis of eye location with respect to the BSL content in the video sequences and calculation of the total and percentage viewing time spent on each region of the BSL video image.

A major contribution of this work was the discovery of a consistent characteristic viewing pattern which demonstrated that the face is the most visually important image region for sign language communication since it is captured in high resolution by the fovea of the eye at the point of gaze. Although the hands of the signer are important for BSL communication, they are not tracked by deaf viewers and so it was concluded that they are processed effectively in low resolution peripheral vision when they are not close enough to the face to be included in the range of the fovea. Knowledge and understanding of the location of gaze during BSL communication contributed to the development of a method which gave priority to that region in the coding scheme.

12.1.3 Novel Methods of Video Compression for BSL Communication Systems

Development of video coding methods for content prioritisation of BSL communication was based on evaluation of available methods identified in Chapter 4, analysis of eye movements during the BSL communication task (Chapter 6), analysis of the content of BSL video material (Chapter 7) and the development and subjective testing of the methods of content prioritisation (spatial and temporal video image foveation) by deaf viewers (Chapters 8 & 9).

Novel video image foveation methods were developed, based on the multi-resolution pyramid model designed by Geisler and Perry (1998), to address the limitations of previous applications of ‘approximate’ foveated imaging methods and methods for segmentation of ‘regions of interest’ along block boundaries (Section 4.2). The new

methods were designed to give priority to the face of the signer which was the region given priority by the foveal response of the HVS in the EMT experiments (Chapter 6).

The potential to improve the perception of BSL video image quality and reduce the number of bits required to encode the content prioritised video was established in analysis of content of the video material and the bit allocations to the different image regions in the encoded BSL material. Video analysis (Chapter 7) revealed that the small optical flows which characterised the face region were given low priority at the expense of large motion flows corresponding to the arms/hands region in standard coding schemes.

The new methods were applied as filters at the pre-processing stage in the CODEC which gave priority to the quality of the face region and produced smooth degradation increasing with distance from the face to the periphery of the image. The methods, spatial image foveation (Section 8.1) and a FWTF (Foveation-Weighted Temporal Filter described in Section 8.2), were designed to allow the strength of foveation to be varied for testing the limits of acceptability in subjective quality testing by deaf users. This enabled the maximum degree (strength) of foveation which would produce acceptable quality for the user to be determined.

12.1.4 Subjective Quality Evaluation Methodology for BSL Video Communication

The application of methods developed to evaluate the quality perceptions of BSL users in this thesis was successful in establishing the maximum strength of spatial and temporal foveation which was acceptable for BSL communication.

The specification and application of experimental design criteria (Chapter 9) contributed to the successful development and refinement of a video quality evaluation method which measured perceived quality for the BSL task rather than the perception of image impairments (overall image fidelity).

The experimental process, which included a review of the developed methods after each application, resulted in the refinement of the methodology to produce a robust method which applied the rigour of ITU recommended procedures (in terms of the

duration and presentation of the stimuli) and reduced the time and effort required by the participants.

A single stimulus method, which presents short BSL video sequences in order of increasing strength of foveation for user evaluation using a five-point criterion-referenced rating scale (Table 9.3) or Binary Acceptability Method (acceptable or unacceptable), is recommended for assessment of user satisfaction for the task (BSL communication) as a result of the experimental work conducted in this thesis.

The development of the subjective quality assessment techniques for BSL video communication contributed to the development of a method of video coding which was acceptable to deaf people.

12.1.5 Performance of the Developed System at Low Bit Rates

Improved performance of the content-prioritised system for BSL video communication was demonstrated in terms of perceived quality for the task (Chapter 10). The novel spatial and temporal video image foveation methods developed in this thesis resulted in improved performance compared with a standard (H.264) coding scheme in terms of perceived quality at low bits rates (below 200 kbps).

The video image foveation methods, applied at the pre-processing stage of the video coding process, achieved bit rate reductions of up to 49% and perceptual gains of at least 50 kbps (at the 100% acceptability level) in subjective quality testing by deaf viewers. The spatial image foveation technique produced the most significant perceptual gains for deaf viewers.

12.2 Original Contributions

The spatial video image foveation method developed for this thesis contributed to the successful development of a real-time application of perceptually optimised video compression in a Proof of Concept project funded by Scottish Enterprise (completed in May 2007). The original contributions of this thesis are described in this section in terms of the development of novel methods and the substantive contributions to knowledge and understanding of systems for BSL video communication.

The presentations, published papers, contributions to patents and awards from this thesis are given in Appendix A.

12.2.1 Methodological Contributions

The user-centred systems design approach in this thesis contributed to the successful development of new methods of subjective quality evaluation and content-prioritised video coding for BSL communication. Methodological contributions were made in the visualisation and analysis of EMT data, analysis of video image content, evaluation of perceived quality (for a specific task) and implementation of video image foveation in the spatial and temporal domain. The techniques and methods developed in this thesis contributed to the development of a system which combined human factors and technical developments to meet the needs of BSL users. In general applications, the approach and techniques developed in this thesis may be applied to the development of optimised systems designed for any specific task or user group.

12.2.2 Substantive Contributions

The main contributions of this work are new insights into visual responses to BSL communication and optimisation of video communication systems based on the limitations and capabilities of the HVS in a 'model of visual perception' (Section 2.5).

The user-centred, task-oriented design approach involved detailed analysis of eye gaze coordinates and frame-by-frame analysis of the image content of BSL video to ensure a strong foundation for the development of a system to meet the specific needs and task requirements of the user.

The establishment of the spatial and temporal visual requirements in the foveal and peripheral fields of vision contributed to knowledge and understanding of the quality requirements of video communication systems for the BSL communication task. This research presented a strong case for the value of eye movement tracking data for system design in an application where there was a clear link between gaze location and the locus of attention in the viewer.

A new approach to video coding optimisation based on visual perception mechanisms is the most significant contribution of this thesis. Research will continue to develop commercial systems to benefit sign language users in the future. The outcomes from this thesis have contributed to this process in a Proof of Concept project, a Patent application and in suggestions for further work (Section 11.3) to take this research forward.

References

ADELSON, E.H. and BERGEN, J.R., 1985. Spatiotemporal energy models for the perception of motion. *Journal of the Optical Society of America,* 2 (2), pp. 284-299

AGRAFIOTIS, D., CANAGARAJAH, N., BULL, D.R. and DYE, M., 2003. Perceptually optimised sign language video coding based on eye tracking analysis. *IEE Electronics Letters,* 39 (24), pp. 1703-1705

AGRAFIOTIS, D., CANAGARAJAH, N., BULL, D.R., KYLE, J., SEERS, H. and DYE M., 2004. A video coding system for sign language communication at low bit rates. *Proceedings of the International Conference on Image Processing (ICIP)*. 24-27 October 2004. Singapore. 1, pp. 441-444

AGRAFIOTIS, D., CANAGARAJAH, N., BULL, D.R., KYLE, J., SEERS, H. and DYE M., 2006. A perceptually optimised video coding system for sign language communication at low bit rates. *Signal Processing: Image Communication*, 21 (7), pp. 531-547

AHMED, N., NATARAJAN T. AND RAO K.R., 1974. Discrete Cosine Transform. *IEEE Transactions on Computers*, C-23 (1), pp. 90-94

ALDRIDGE, R., DAVIDOFF, J., GHANBARI, M., HANDS, D. and PEARSON, D., 1995. Measurement of scene-dependant quality variations in digitally coded television pictures. *IEE Proceedings of Vision, Image and Signal Processing*, 142 (3), pp.149-154

ARIFF, G., DONCHIN, O., NANAYAKKARA, T and SHADMEHR, R., 2002. A real-time state predictor in motor control: Study of saccadic eye movements during unseen reaching movements. *Journal of Neuroscience*, 22 (17), pp. 7721-7729

ARRINGTON RESEARCH, 2002. ViewPoint Eye Tracker^{TM} PC-60 Software Users Guide. Version 2.5. 3 July 2002. Arrington Research Inc.

BAKER, C.L. and BRADDICK, O.J., 1985. Eccentricity-dependent scaling of the limits of short-range motion perception. *Vision Research,* 25 (6), pp. 803-12

BARRON, J.L., FLEET, D.J. and BEAUCHEMIN, S.S., 1994. Performance of optical flow techniques. *International Journal of Computer Vision,* 12 (1), pp. 43-77

BBC, 2002. Signing avatar. *BBC World*. [online] London: British Broadcasting Corporation. Available from: http://www.bbcworld.com (archive 29 August 2002) [Accessed 9 June 2006].

BOUCH, A., WATSON, A. and SASSE, M.A., 1998. [Poster]. QUASS – A tool for measuring the subjective quality of real-time multimedia audio and video. *Human Computer Interaction (HCI) Conference*, 1-4 September 1998, Sheffield, England.

BRUCE, V., GREEN, P.R. and GEORGESON, M.A., 2003. *Visual perception: Physiology, psychology and ecology*, 4th ed. Hove & London: Psychology Press.

BURR, D.C. and ROSS, J., 1986. Visual processing of motion. *Trends in Neuroscience*, 9 (7), pp. 304-7

BUSWELL, G.T., 1935. How people look at pictures: A study of the psychology of perception in art. In: S.E. PALMER, ed. *Vision science: photons to phenomenology*. Cambridge, Massachusetts: MIT Press.

CACDP, 2003. Curriculum for Level 2 certificate in British Sign Language. Durham: Council for the Advancement of Communication with Deaf People (CACDP).

CAVANAGH, P., 1992. Attention-based motion perception. *Science*, 257 (5076), pp.1563-1565

CAVANAGH, P. and MATHER, G., 1989. Motion: The long and short of it. *Spatial Vision*, 4, pp.103-129

CCITT, Recommendation H.120, 1989. *CODECs for videoconferencing using primary digital group transmission*. Geneva.

CCITT, Recommendation H.261, 1990. *Video CODEC for audio-visual services at px64kbits/s*. Geneva.

CHOI, C.S. and TAKEBE, T., 1994. Analysis and synthesis of facial image sequences in model-based image coding. *IEEE Transactions Video Technology*, 4 (3), pp. 257-275

CHUA, H.F., BOLAND, J.E. and NISBETT, R.E., 2005. Cultural variation in eye movements during scene perception. *Proceedings of the National Society of Science*, 102 (35) pp. 12629-12633

CHUBB, C. and SPERLING, G., 1988. Drift-balanced random dot stimuli: A general basis for studying non-Fourier motion perception. *Journal of the Optical Society of America*, A, 5, pp. 1986-2007

COIMBRA, M.T. and DAVIES, M., 2005. Approximating optical flow within the MPEG-2 compressed domain. *IEEE Transactions on Circuits and Systems for Video Technology,* 15 (1), pp. 103-107

COMMUNICATION SERVICE FOR THE DEAF (CSD), 2005. *BDA-CSA launches sign language video relay service*. [online] 23 August 2005.
Available from: http://www.c-s-d.org [Accessed 12 June 2006].

CORBETTA, M., AKBUDAK, E., CONTURO, T.E., SNYDER, A.Z., OLLINGER, J.M., DRURY, H.A., LINENWEBER, M.R., PETERSEN, S.E., RAICHLE, M.E., VAN ESSEN, D.C. and SHULMAN, G.L., 1998. A common network of functional areas for attention and eye movements. *Neuron*, 21, pp. 767-773

CORINA, D., BELLUGI, U. and REILLY, J., 1999. Neuropsychological studies of linguistic and affective facial expressions in deaf signers. *Language and Speech,* 42 (2-3), pp. 307-331

CORINA, D.P., 1989. Recognition of affective and non-canonical linguistic facial expressions in hearing and deaf subjects. *Brain and Cognition*, 9, pp. 227-237

CSIRO, 2001. *Image Motion Analysis Research*. The Image Analysis Group, CSIRO (Commonwealth Scientific and Industrial Research Organisation) Division of Mathematical and Information Sciences. [online] Available from: http://www.cmis.csiro.au/IAP [Accessed 26 November 2001].

CSIRO, 2004. *Image Motion Analysis Research*. The Image Analysis Group, CSIRO (Commonwealth Scientific and Industrial Research Organisation) Division of Mathematical and Information Sciences. [online] Available from: http://www.cmis.csiro.au/IAP [Accessed 3 October 2006].

CUMMING, G.D., 1978. Eye movements and visual perception. In E.C. CARTERETTE AND M.P. FRIEDMAN, eds. *Handbook of perception*. Cambridge, Massachusetts: Academic Press. pp. 221-255

DUCHOWSKI, A.T., 2002. A breadth-first survey of eye tracking applications. *Behaviour Research Methods, Instruments, & Computers*, 34 (4), pp. 455-470

DUCHOWSKI, A.T., 2003. *Eye tracking methodology theory and practice*. London: Springer-Verlag.

DUGÉNIE, P., MUNRO, A.T. and BARTON, M.H., 2002. Toward assessing subjective quality of service of conversational mobile multimedia applications delivered over the Internet: A methodological study. *IEEE Transactions on Multimedia,* 4 (1), pp. 59-67

ELEFTHERIADIS, A. and JACQUIN, A., 1995. Automatic face location, detection and tracking for model-assisted coding of video teleconferencing sequences at low bit rates. *Signal Processing: Image Communication,* 7 (3), pp. 231-248

EMMORY, K., MCCULLOUGH, S. and BRENTARI, D., 2003. Categorical perception in ASL. *Language and Cognitive Processes,* 18 (1), pp. 21-45

FELS, D., RICHARDS, J., HARDMAN, J., SOUDAIN, S. and SILVERMAN, C., 2004. American Sign Language of the Web. *Proceedings of ACM Computer Human Interaction*, 24-29 April 2004, Vienna, Austria.

FINDLAY, J.M. and GILCHRIST, I.D., 2003. *Active vision: The psychology of looking and seeing*. New York: Oxford University Press.

FINNEY, E.M., FINE, I. and DOBKINS, K.R., 2001. Visual stimuli activate auditory cortex in the deaf. [online] *Nature Neuroscience*, 12 November 2001. Available from: http://neurosci.nature.com [Accessed 25 April 2004]

FIORENTINI, A., 1989. Differences between fovea and parafovea in visual search processes. *Vision Research,* 29 (9), pp. 1153-1164

FLEET, D.J. and JEPSON, A.D., 1990. Computation of component image velocity from local phase information. *International Journal of Computer Vision,* 5 (1), pp. 77-104

GALE, A.G., 1997. Human response to visual stimuli. In: W. R. Hendee and P. N. T. Wells, eds. *The Perception of Visual Information.* 2nd ed. New York: Springer-Verlag.

GALVIN, B., MCCANE, B., NOVINS, K., MASON, D. and MILLS, S., 1998. Recovering motion fields: An evaluation of eight optical flow algorithms. *Proceedings of the Ninth British Machine Vision Conference*, 14-17 September 1998, Southampton, UK. pp. 195-204

GAUTAMA, T. and VAN HULLE, M.M., 2002. A phase-based approach to the estimation of the optical flow fields using spatial filtering. *IEEE Transactions on Neural Networks,* 13 (5), pp. 1127-1136

GAZZANIGA, M.S., IVRY, R.B. and MANGUN, G.R., 1998. *Cognitive neuroscience: The biology of the mind*. New York: Norton.

GEISLER, W.S. and PERRY, J.S., 1998. A real-time foveated multi-resolution system for low bandwidth video communication. In: B. Rogowitz and T. Pappas, eds. *Proceedings of SPIE Conference on Human Vision and Electronic Imaging*, 20-30 January 1998, San Jose, CA. 3299, pp. 294-305

GEISLER, W.S. and PERRY, J.S., 1999. Variable resolution displays for visual communication and simulation. *The Society for Information Display*, 30, pp. 420-423

GEISLER, W.S. and PERRY, J.S., 2002. Gaze contingent real-time simulation of arbitrary visual fields. *Proceedings of SPIE Conference on Human Vision and Electronic Imaging*, 21 – 24 January 2002, San Jose, CA. 4662, pp. 57-69

GEORGESON, M.A. and HAMMETT, S.T., 2002. Seeing Blur: 'Motion Sharpening' without Motion. *Proceedings of the Royal Society of London*, 269, pp. 1429-1434

GINSBURG, A.P. and HENDEE, W.R., 1997. Quantification of visual capability. In: W.R. Hendee, and P.N.T. Wells, eds. *The Perception of Visual Information*. 2nd ed. New York: Springer-Verlag.

GIROD, B., 1992. Psychovisual aspects of image communication. *Signal Processing*, 28 (3), pp. 239-251

GOLDBERG, J.H., STIMSON, M.J., LEWENSTEIN, M., SCOTT, N. and WICHANSKY, A.M., 2002. Eye tracking in web search tasks: Design implications. *Proceedings of the ACM Symposium on Eye Tracking Research and Applications*, New Orleans, Louisiana. ACM Press: New York, USA. pp. 51-58

GRAYSON, D.M. and MONK, A.F., 2003. Are you looking at me? Eye contact and desktop video conferencing. *ACM Transactions on Human Computer Interaction*, 10 (3), pp. 221-243

GRONER, R. and GRONER, M.T., 1989. Attention and eye movement control: An overview. *European Archives of Psychiatry and Clinical Neuroscience*, 239 (1), pp. 9-16

GUBA, E., WOLF, W. DEGROOT, S., KNEMEYER, M., VAN ATTA, R. and LIGHT, L. 1964. Eye movements and TV viewing in children. *Audio-Visual Communication Review*, 12, pp. 386-410

GUO, K., MAHMOODI, S., ROBERTSON, R.G. and YOUNG, M.P., 2006. Longer fixation time while viewing face images. *Experimental Brain Research*, 171, pp. 91-98

HELLSTRÖM, G., 1997. Quality Measurement on Video Communication for Sign Language, *Proceedings of the Sixteenth International Symposium on Human Factors in Telecommunications*, 12-16 May 1997, Oslo, Norway. pp. 217-224

HENDEE, W.R. and WELLS, P.N.T., eds. 1997. *The perception of visual information*. 2nd ed. New York: Springer-Verlag.

HENDERSON, J.M. and HOLLINGWORTH, A., 1999. High level scene perception. *Annual Review of Psychology*, 50, pp. 243-271

HILDRETH, E.C., 1984. Measurement of visual motion. Cambridge, Massachusetts: MIT Press.

HIRSCH, J. and CURCIO, C.A., 1989. The spatial resolution capacity of human foveal retina. *Vision Research*, 29, pp. 1095-1101

HOFFMAN, J. E. and SUBRAMANIAM, B., 1995. The role of visual attention in saccadic eye movements. *Perception and Psychophysics*, 57 (6), pp. 787-795

HOLLANDER, M., and WOLFE, D.A., 1999. *Nonparametric Statistical Methods.* New York: Wiley.

HORN, B.K.P. and SCHUNCK, B.G., 1981. Determining optical flow. *Artificial Intelligence* 17, pp. 185-203

HYÖNÄ, J. and LORCH, R.F., 2004. Effects of topic headings on text processing: Evidence from adult readers' eye fixation patterns. *Learning and Instruction*, 14, pp. 131-152

IKEDA, M. and TAKEUCHI, T., 1975. Influence of foveal load on the functional visual field. *Perceptual Psychophysics,* 102 (2), pp. 224-249

ISO/IEC, 1993. 11172-2: Information Technology - Coding of moving pictures and associated audio for digital storage media at up to about 1.5Mbits/s – Part 2: Video. [MPEG-1 video].

ISO/IEC, 1995. 13818-2: Information Technology - Generic coding of moving pictures and associated audio information: Video. [MPEG-2 video].

ISO/IEC, 1998. 14996-2: Information Technology - Coding of audio-visual objects Part 2: Visual. [MPEG-4 visual].

INTERNATIONAL TELECOMMUNICATIONS UNION, 2006. *The International Telecommunications Union*. [Online] Available from: http://www.itu.int/home/index.html [Accessed 23 October 2006].

ITU-T, 1993. *Recommendation H.261: Video codec for audio-visual services at px64kbits/s*, version 2. Geneva: ITU-T.

ITU-T, 1995. *Recommendation H.263: Video Coding for Low Bit Rate Communication*, version 1. Geneva: ITU-T.

ITU-T, 1998. *Recommendation H.263: Video Coding for Low Bit Rate Communication*, version 2. Geneva: ITU-T.

ITU-T, 1999. *Recommendation P.910: Subjective video quality assessment methods for Multimedia Applications*. September, 1999. Geneva: International Telecommunications Union Standardisation Sector (ITU-T).

ITU-T, 2000. *Recommendation H.263: Video Coding for Low Bit Rate Communication*, version 3. Geneva: ITU-T.

ITU-T, 2002. *Recommendation BT.500-11: Methodology for the subjective assessment of the quality of television pictures*. June. Geneva: International Telecommunications Union Standardisation Sector (ITU-T).

ITU-T, 2003. *Recommendation H.264 (05/03): Advanced video coding for generic audiovisual services*. Geneva: ITU-T. (superseded)

ITU-T, 2005. *Recommendation H.264 (03/05): Advanced video coding for generic audiovisual services*. Geneva: ITU-T.

ITU-T Recommendation H.264 & ISO/IEC 14496-10, 2005. Advanced Video Coding for Generic Audiovisual Services (version 3). [Scalable Video Coding].

ITU-T STUDY GROUP 9, 2004. Question 14/9: Objective and subjective methods for evaluating perceptual audiovisual quality in multimedia services. [online] Available from: http://www.itu.int/ITU-T/studygroups/com09/sg9-q14.html [Accessed 2 November 2006].

ITU-T STUDY GROUP 16, 1998. *Draft application profile: Sign language and lip reading real time conversation usage of low bit rate video communication*. Geneva: ITU-T.

JACOB, R.J.K., 1991. The use of eye movements in human-computer interaction techniques: What you look at is what you get. *ACM Transactions on Information Systems,* 9 (3), pp. 152-169

JACOB, R.J.K., 1993. Eye-movement-based human-computer interaction techniques: Toward non-command interfaces. In: H. R. HARTSON and D. HIX, eds. *Advances in Human-Computer Interaction*. New Jersey: Ablex. pp. 151-190

JACOB, R. J. K., 1994. New human-computer interaction techniques. In: M. BROUWER-JANSE and T. HARRINGTON, eds. *Human-Machine Communication for Educational Systems Design*. Berlin: Springer-Verlag. pp. 131-138

JACOB, R.J.K., 1995. Eye tracking in advanced interface design. In: W. BARFIELD and T. FURNESS, eds. *Advanced Interface Design and Virtual Environments*. Oxford: Oxford University Press. pp. 258-288

JACOB, R.J.K. and KARN, K.S., 2003. Eye Tracking in Human-Computer Interaction and Usability Research: Ready to Deliver the Promises. In: R. RADACH, J. HYÖNÄ, and H. DEUBEL, eds. *The mind's eye: Cognitive and applied aspects of eye movement research*. Boston: North-Holland/Elsevier. pp. 573-605

JONES, B.L. and MCMANUS, P.R., 1986. Graphic scaling of qualitative terms. Journal of the Society of Motion Picture and Television Engineers. November 1986. pp. 1166-1171

JULESZ, B., 1986. A brief outline of the texton theory of human vision. *Trends in Neuroscience*, 7, pp. 41-45

JVT [20] , 2005. H.264/MPEG4-Part 10/AVC: Advanced video coding for generic audiovisual services. Geneva: ITU-T.

[20] JVT (Joint Video Team) comprising the ITU-T VCEG (Video Coding Experts Group) and ISO/IEC MPEG (Moving Picture Experts Group)

KELSEY, C.A., 1997. Detection of Visual Information. In: W. R. HENDEE and P. N. T. WELLS, eds. *The Perception of Visual Information*. New York: Springer-Verlag.

KLEINFELDER, K., 1999. Foveated imaging on a smart focal plane. FOVIS, March 12. [on-line]. Available from: http://ise.stanford.edu/class/psych221/projects/99/stuartk/fovis.html [Accessed 15 January 2004].

LAND, M.F., MENNIE, N. and RUSTED, J., 1999. The roles of vision and eye movements in the control of activities of everyday living, *Perception*, 28, pp. 1311-28

LEDGEWAY, T. and SMITH, A.T., 1994. The duration of the motion aftereffect following adaptation to first- and second-order motion. *Perception*, 23, pp. 1211-1219

LEE, C., KWON, O., JEONG, T., CHO, S. and KIM, H., 2002. Weighted PSNR for objective measurement of video quality. *Proceedings of Visualisation, Imaging, and Image Processing*. 9-12 September 2002, Marbella, Spain.

LOFTUS, G.R., 1985. Picture perception: Effect of luminance level on available information and information extraction rate. *Journal of Experimental Psychology: Human Perception and Performance*, 4, pp. 565-572

LUCAS, B.D. and KANADE, T., 1981. An interactive image registration technique with an application to stereo vision. *Proceedings of the Seventh International Joint Conference of Artificial Intelligence (IJCAI),* pp. 674-679

MACK, A. and ROCK, I., 1998. Inattentional blindness. Cambridge, Massachusetts: MIT Press.

MARPE, D., WIEGAND, T. and SULLIVAN, G.J., 2006. The H.264/MPEG4-AVC Standard and its Fidelity Range Extensions, *IEEE Communications Magazine*, October 2006.

MANNAN, S.K., RUDDOCK, K.H. and WOODING, D.S., 1995. Automatic control of saccadic eye movements made in visual inspection of briefly presented 2-D images. *Spatial Vision*, 9, pp. 363-386

MCCARTHY, J.D., SASSE, M.A. and MIRAS, D., 2004. Sharp or smooth? Comparing the effects of quantisation vs. frame rate for streamed video. *Proceedings of Computer Human Interaction (CHI),* 24-29 April 2004, Vienna, Austria.

MCCAUL, T., 1997. Video-based telecommunications technology and the deaf community. *Report of Australian Communication Exchange*. Queensland: Australian Communication Exchange Ltd.

MORLAND, A.B, JONES, S.R., FINLAY, A.L, DEYZAC, E., LE, S. and KEMP, S., 1999. Visual perception of motion, luminance and colour in a human hemianope. *Brain,* 122, pp. 1183-1198

MUIR, L.J. and RICHARDSON, I.E.G., 2002. Video telephony for the deaf: Analysis and development of an optimised video compression product. *Proceedings of the Tenth ACM International Multimedia Conference*, 1-6 December 2002, Juan-Les-Pins, France. pp. 650-652

MUIR, L.J. and RICHARDSON, I.E.G., 2003. Video coding for sign language communication. [presentation] *'Silent Progress' ESRC Seminar*, 28 February -1 March 2003, University of Bristol.

MUIR, L.J. and RICHARDSON, I. E. G., 2005. Perceptions of sign language and its application to visual communications for deaf people. *Journal of Deaf Studies and Deaf Education*, *10* (4), pp. 390-401

MUIR, L.J., RICHARDSON, I.E.G. and HAMILTON, K., 2005. Visual perception of content-prioritised sign language video quality. *Proceedings of the IEE International Conference on Visual Information Engineering*, 4-6 April 2005, University of Glasgow, Glasgow. pp.17-22

MUIR, L.J., RICHARDSON, I.E.G. and LEAPER, S., 2003. Gaze tracking and its application to video coding for sign language. *Proceedings of the Twenty-third International Picture Coding Symposium*, 22-26 April 2003, St. Malo, France. pp. 321-325

MULLIN, J., JACKSON, M., HENDERSON, A.H., SMALLWOOD, L., SASSE, M.A., WATSON A. and WILSON, G., 2002. Assessment Methods for Assessing Video Quality in Real-time

Interactive Communications. *Report of the ETNA (Evaluation Technology for Networked multimedia Applications) Project* [online] Available from: http://www-mice.cs.ucl.ac.uk/multimedia/projects/etna [Accessed 21 October 2004].

NIELSEN, J., 2006. F-Shaped Pattern for Reading Web Content. *Alertbox*, April 17. [online] Available from: http://www.useit.com/alertbox/reading_pattern.html [Accessed 19 September 2006].

NISHIDA, S., LEDGEWAY, T. and EDWARDS, M., 1997. Dual multiple-scale processing for motion in the human visual system. *Vision Research*, 37, pp. 2685-2698

NOTON, D. and STARK, L., 1971. Scanpaths in saccadic eye movements while viewing and recognising patterns. *Vision Research,* 11, pp. 929-942. In: J. M. FINDLAY, and I. D. GILCHRIST, 2003. *Active vision: The psychology of looking and seeing*. Oxford: Oxford University Press.

O'MALLEY, C., BROOKS, P., BRUNDELL, P., HAMNES, K., HEIESTAD, S., HEIM, J., HESTNES, B., HEYDARI, B., SCHLIEMANN, T., SKJETNE, J.H. and ULSETH, T., 1999. Results of video telephony experiments. ACTS Project AC314 Vis-à-vis: Fitness-for-purpose of videotelephony in face-to-face situations. *CEC Deliverable A314/UoN/SoP/DR/P003/b1*, 30 June 1999.

O'REGAN, J.K., 1992. Solving the "real" mysteries of visual perception: The world as an outside memory. *Canadian Journal of Psychology,* 46 (3), pp. 461-488

OSTERBERG, G., 1935. Topography of the layer of rods and cones in the human retina. *Acta Ophthalmologica*, 13 (6), pp. 1-103

PALMER, S.E., 2002. *Vision science: Photons to phenomenology*. Cambridge, Massachusetts: MIT Press.

PORIKLI, F.M., 2004. Real-time video object segmentation for MPEG encoded video sequences. *Proceedings of the SPIE Conference on Real Time Imaging VIII*, 5297, pp. 195-203

POSNER, M.I., 1980. Orienting of attention. *Quarterly Journal of Experimental Psychology*, 32, pp. 3-25. In: K. RAYNER, 1998. Eye movements in reading and information processing: 20 years of research. *Psychological Bulletin,* 124 (3), pp. 372-422

RAFF, U., 1997. Visual data formatting. In: W.R. HENDEE, AND P.N.T. WELLS, eds. *The Perception of Visual Information*. 2nd ed. New York: Springer-Verlag.

RAO, K.R. and YIP, P., 1990. *Discrete Cosine Transform: Algorithms, advantages, applications*. San Diego, CA: Academic Press.

RAPANTZIKOS, K. and ZERVAKIS, M., 2005. Robust optical flow estimation in MPEG sequences. *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, 18-23 March 2005, Philadelphia.

RAYNER, K., 1998. Eye movements in reading and information processing: 20 years of research. *Psychological Bulletin*, 124 (3), pp. 372-422

REICHARDT, W., 1969. Movement perception in insects. In: W. Reichardt, ed. *Processing of optical data by organisms and by machines*. New York: Academic Press.

REKLEITIS, I.M., 1996. Steerable filters and cepstral analysis for optical flow calculation from a single blurred image. *Vision Interface*, 12 (1), pp.159-166

RICHARDSON, I.E.G., 2003. *H.264 and MPEG-4 video compression*, Chichester: John Wiley & Sons.

RICHARDSON, I.E.G., 2002. *Video codec design: Developing image and video compression systems*, Chichester: John Wiley & Sons.

RICHARDSON, I.E.G. and KANNANGARA, S., 2004. Fast subjective video quality measurement with user feedback, *Electronics Letters*, 40 (13), pp. 799-800

RICHARDSON, I.E.G., ZHAO, Y and MUIR, L.J., 2002. Implementing MPEG-4 Visual in software. *IEE Seminar on Visual Media Standards*, 25 April 2002, London, UK.

RIEGELSBERGER, J., SASSE, M.A. and MCCARTHY, J., 2002. Eye catcher or blind spot? The effect of photographs of faces on E-commerce sites. *Proceedings of the second IFIP conference on e-commerce, e-business, e-government (i3e)*, October 7-9 2002, Lisbon, Portugal, pp. 383 – 398

ROBINSON, D., 1968. The oculomotor control system: A review. *Proceedings of the IEEE,* 56 (6), pp.1032-1047

ROBSON, J.G., and GRAHAM, N., 1981. Probability summation and regional variation in contrast sensitivity across the visual field. *Vision Research*, 21, pp. 409-418

RNID (Royal National Institute for the Deaf), 2004. The Royal National Institute for the Deaf. *RNID Newsletter.* March 2004. London: Royal National Institute for the Deaf.

RNID (Royal National Institute for the Deaf), 2006a. RNID Sign Talk. *RNID Newsletter.* 23 March 2006. London: Royal National Institute for the Deaf.

RNID (Royal National Institute for the Deaf), 2006b. *Virtual Signing*. [online] London: Royal National Institute for the Deaf. Available from: http://www.rnid.org.uk (How we help: Research and Technology: Communication and Broadcasting) [Accessed 9 June 2006].

SADKA, A.H., 2002. *Compressed video communications*, Chichester: John Wiley & Sons.

SALVUCCI, D.D. and GOLDBERG, J.H., 2000. Identifying fixations and saccades in eye tracking protocols. *Proceedings of ACM Eye Tracking Research and Applications Symposium*. New York, USA: ACM Press. pp 71-78

SAXE, D.M. and FOULDS R.A., 2002. Robust region of interest coding for improved sign language telecommunication, *IEEE Transactions on Information Technology in Biomedicine*, 6 (4), pp. 310-316

SCHIERL, T., GÄNGER, K., HELLGE, C., WIEGAND, T and STOCKHAMMER, T., 2006. SVC-based multisource streaming for robust video transmission in mobile and ad hoc networks. *IEEE Wireless Communications*, October 2006, pp 96 - 103

SCHNIPKE, S.K. and TODD, M.W., 2000. Trials and tribulations of using an eyetracking system. *Proceedings of the ACM Computer-Human Interaction Conference on Human Factors in Computing Systems*, The Hague, Netherlands. New York: ACM Press. pp. 273-274

SCHUMEYER, R., HEREDIA, E., and BARNER, K., 1997. Region of interest priority coding for sign language videoconferencing, *Proceedings of the First IEEE Workshop on Multimedia Signal Processing*, June 1997, Princeton, New Jersey. pp. 531-536

SCOTTISH EXECUTIVE, 2004. *Written answers: Sign Language*. 14 January 2004.

SHARP, P. and PHILIPS, R., 1997. Physiological optics. In: W.R. HENDEE and P.N.T. WELLS, eds. *The Perception of visual information*. New York: Springer-Verlag.

SHEIKH, H.R., EVANS, B. L. and BOVIK, A.C., 2003. Real-time foveation techniques for low bit rate video coding. *Real Time Imaging*, 9 (1), pp. 27-40

SHEPHERD, M., FINDLAY, J.M. and HOCKEY, R.J., 1986. The relationship between eye movements and spatial attention. *Quarterly Journal of Experimental Psychology*, 38A, pp. 475-491

SHIORI, S. and IKEDA M., 1989. Useful resolution for picture perception as a function of eccentricity. *Perception*, 18, pp. 347-361

SIKORA, T., 1999. Digital coding standards and their role in video communications. In: J.S. BYRNES, ed. *Signal Processing for Multimedia*. Amsterdam: IOS Press.

SILVERSTEIN, D.A. and FARRELL, J.E., 1996. The relationship between image fidelity and image quality. *Proceedings of the IEEE International Conference on Image Processing*, 1, pp. 881-884

SIMPKINS, J. and WILLIAMS, J.I., 1988. *Advanced human biology*. London: Unwin Hyman.

SIPLE, P., 1978. Visual constraints for sign language communication. *Sign Language Studies,* 19, pp. 95-110

SORYANI, M. and CLARKE, R.J., 1992. Segmented coding of digital image sequences, *Communications, Speech and Vision, IEE Proceedings I*, 139 (2), pp. 212 – 218

SPILLER, F., 2004. Eye tracking usability studies – Usability Holy Grail? *Demystifying Usability*, [online] Available from: http://experiencedynamics.blogs.com/site_search_usability/2004/12/eyetracking_stu.html [Accessed 14 November 2006].

SPILLER, F., 2005. How many users should you test with in usability testing? *Demystifying Usability*, [online] Available from: http://experiencedynamics.blogs.com/site_search_usability/2005/01/latest_research.html [Accessed 14 November 2006].

STOKOE, W.C., 2001. The study and use of sign language. *Sign Language Studies,* 1 (4) pp. 369-406

STONER, G.R., ALBRIGHT, T.D. and RAMACHANDRAN, V.S., 1990. Transparency and coherence in human motion perception. *Nature*, 344 (6262), pp. 153-5

THOMPSON, R. and EMMORY, K., 2003. The relationship of eyegaze and agreement morphology in ASL: An eye tracking study. *Meeting of the Linguistic Society of America*. January 2003, Atlanta, GA.

VAN DIEPEN, P.M.J., WAMPERS, M. and D'YDEWALLE, G., 1998. Functional division of the visual field: Moving masks and moving windows. In: G. Underwood, ed. *Eye guidance in reading and scene perception*. Amsterdam: Elsevier. pp. 337-355

van Santen, J.P.H. and Sperling, G., 1985. Elaborated Reichardt detectors. *Journal of the Optical Society of America*, A 2 (2), pp. 300-321

VERTANEN, M.T., GLEISS, N. and GOLDSTEIN, M., 1995. On the use of evaluative category scales in telecommunications. *Proceedings of Human Factors in Telecommunications*, pp. 253-260

VIVIANNI, P., 1990. Eye movements in visual search: Cognitive, perceptual and motor control aspects. In: E. Kowler, ed. *Eye Movements and Their Role in Visual and Cognitive Processes*, Amsterdam: Elsevier. pp. 353-93

WANG, Z., BOVIK, A.C., SHEIKH, H.R. and SIMONCELLI, E.P., 2004. Image quality assessment: From error visibility to structural similarity, *IEEE Transactions on Image Processing*, 13 (4), pp. 600-612

WANG, Z., LU, L. and BOVIK, A.C., 2003. Foveation scalable video coding with automatic fixation selection. *IEEE Transactions on Image Processing*, 12 (2), pp. 243-254

WATSON, A., 2001. *Assessing the quality of audio and video components in desktop multimedia conferencing*. PhD thesis, University of London.

WEBB, N. and RENSHAW, T., 2005. Commercial uses of eye tracking. [Workshop] *Nineteenth British Human Computer Interaction Group Annual Conference*, 5-9 September 2005, Napier University, Edinburgh.

WEDI, T. and KASHIWAGI, Y., 2004. Subjective quality evaluation of H.264/AVC FRExt for HD movie content. Joint Video Team document JVT-LO33, July 2004.

WESTLAND, S., OWENS, H., CHEUNG, V. and PATERSON-STEPHENS, I., 2006. Model of luminance contrast-sensitivity function for application to image assessment. *Color Research and Application*, 31 (4), pp. 315-319

WIEGAND, T., SCHWARZ, H., JOCH, A., KOSSENTINI, F. and SULLIVAN, G.J., 2003. Rate-constrained coder control and comparison of video coding standards. *IEEE Transactions on Circuits and Systems for Video Technology*, 13 (7), pp. 688-703

WIEGAND, T., SULLIVAN, G.J., BJONTEGAARD, G., and LUTHRA, A., 2003. Overview of the H.264/AVC video coding standard. *IEEE Transactions on Circuits and Systems for Video Technology*, 13 (7), pp. 560-576

WILSON, G. and SASSE, M.A., 2000. Do users always know what's good for them? Utilising physiological responses to assess media quality. *Proceedings of the*

Fourteenth Annual Conference of the British HCI Group. 5-8 September 2000, Sunderland, UK. pp 327-339

WINKLER, S., 2005. *Digital Video Quality: Vision Models and Metrics*, Chichester: John Wiley & Sons.

WOELDERS, W.W., FROWEIN, H.W., NIELSEN, J., QUESTA, P. and SANDINI, G., 1997. New developments in low-bit rate videotelephony for people who are deaf. *Journal of Speech, Language and Hearing Research,* 40, pp. 1425-1433

YARBUS, A.L., 1967. Eye movements and vision. In: S.E. Palmer, ed. *Vision science: Photons to phenomenology*. Cambridge, Massachusetts: MIT Press.

ZAGIER, E.J.S., 2001. A human's eye view: Motion and frameless rendering. *ACM Crossroads*. [online]. Available from: http://www.acm.org/crossroads/xrds3-4/ellen.html [Accessed 15 January 2004]

ZHONG, Y., RICHARDSON, I.E.G., SAHRAIE, A. and MCGEORGE, P., 2003. Qualitative and quantitative assessment in video compression. *Twelfth European Conference on Eye Movements*, 20-24 August 2003, Dundee, Scotland.

Bibliography

AMERICAN PSYCHOLOGICAL ASSOCIATION (APA), 2001. *Publication manual of the American Psychological Association*. 5th ed. Washington: APA.

CHEN, C. and HSU, C., 2001. Content-based hybrid DPCM/Classified vector quantisation for coding video telephony sequences. *Journal of Visual Communication and Image Representation*, 12, pp. 152-168

COLLEWIJN, H.J., 1998. Eye movement recording. In: R. H. S. Carpenter and J. G. Robson eds. *Vision research: A practical guide to laboratory methods*. Oxford: Oxford University Press. pp 245-285

CONNOR, S., 2004. Blind stroke victim learns to 'see' facial expressions. *Independent News*, 15 December 2004.

FISCHER, S.D., DELHORNE, L.A. and REED, C.M., 1999. Effects of rate presentation on the reception of American Sign Language. *Journal of Speech Language and Hearing Research*, 42, pp. 568-582

GANONG, W.F., 2001. *Review of medical physiology* (20th ed). New York: Lange Medical Books/McGraw Hill Medical Publishing Company.

HENDERSON, J.M. AND HOLLINGWORTH, A., 2003. Eye movements and visual memory: Detecting changes to saccade targets in scenes. *Perception and Psychophysics*, 65 (1), pp. 58 – 71

HICKOK, G., BELLUGI, U. and SKIMA, E., 2001. How does the human brain process language? New studies of deaf signers hint an answer. *Scientific American*, June 2001, pp 58-65

JARVIS, J.R., TAYLOR, N.R., PRESCOTT, N.B., MEEKS, I. and WATHES, C.M., 2002. Measuring and modelling the photopic flicker sensitivity of the chicken (Gallus g. domesticus). *Vision Research*, 42, pp. 99-106

KAMINATI, Y and TONG, F., 2005. Mindreading machine knows what you see. *New Scientist*, 25 April, 2005. [online]. Available from: http://www.newscientist.com/article/dn7304.html [Accessed 25th April 2005].

MOLER, C., 2002. *Numerical Computing with Matlab*. Philadelphia: Society for Industrial and Applied Mathematics (SIAM). Companion site: http://www.mathworks.com/moler

SCHWARTZ, E.L., 1980. Computational Anatomy and Functional Architecture of Striate Cortex: A Spatial Mapping Approach to Perceptual Coding. *Vision Research* 20, pp. 645-669

SNOWDEN, R., THOMPSON, P. and TROSCIANKO, T., 2006. *Basic Vision: An Introduction to Visual Perception*. Oxford: Oxford University Press.

WANG, Z and BOVIK, A.C., 2005. Foveated Image and Video Coding. In: H. R. Wu and K. R. Rao. eds. *Digital Video Image Quality and Perceptual Coding*, Marcel Dekker Series in Signal Processing and Communications.

WATSON, A.B. (ed), 1993. *Digital Images and human vision*. MA:MIT Press.

WATSON, A.B., 1994. Image compression using the Discrete Cosine Transform. *Mathematica Journal*, 4(1), pp 81-88

Appendices

Appendix A: List of Publications

Seminar Presentations

RICHARDSON, I. E. G., ZHAO, Y AND MUIR, L.J., 2002. Implementing MPEG-4 Visual in software. *IEE Seminar on Visual Media Standards*, London, UK, April 25 2002.

Notes: An oral presentation was given at the IEE seminar by each of the authors on their work to optimise video coding applications. The aim, background and optimisation techniques proposed for optimising video compression standards for sign language communication was introduced in the context of development within the MPEG-4/H.264 Standard.

MUIR, L. J. AND RICHARDSON, I.E.G., 2003. Video coding for sign language communication. [presentation] *'Silent Progress' ESRC Seminar*, University of Bristol, 28 February -1 March 2003.

Notes: The authors were invited to contribute to this ESRC funded seminar at the University of Bristol which brought academic researchers, technology providers and representatives of the deaf community together for knowledge sharing and to direct future research in this field.

Abstract: Video telephony offers potential for direct personal communication of sign language at a distance. Current standards and systems do not provide sufficient quality for freely expressed sign language and finger spelling over low bit rate channels. The aim of this research is to develop an optimised video compression solution for sign language communication. The objectives are to identify important regions of sign language video sequence images for deaf people, develop a video coding solution and apply an appropriate quality assessment methodology. Results of eye movement tracking experiments with deaf volunteers indicate that the most important region of sign language video content is the face, with some vertical excursions and with hand movements mainly observed in peripheral vision. Based on the outcome of these experiments it is proposed to apply content-based prioritisation to video coding using standards such as MPEG-4 Visual and/or H.264.

Conference Papers

MUIR, L. J. AND RICHARDSON, I.E.G., 2002. Video telephony for the deaf: Analysis and development of an optimised video compression product. *Proceedings of the tenth ACM International Multimedia Conference*, Juan-Les-Pins, France, 1-6 December 2002, pp 650-652

Notes: ACM Multimedia is an annual international conference, covering all aspects of multimedia computing: from underlying technologies to applications, theory to practice, and servers to networks to devices. 'World first' ground breaking new research was presented, at the 2002 conference in France, on video communication systems for deaf people based on eye movement tracking research.

Abstract: The multimedia capability of video telephony and video conferencing systems has many applications and benefits. This paper describes research and development that aims to optimise video compression systems for a specific application - personal communication at a distance for deaf people. Results of eye movement tracking experiments and proposals for image content prioritisation based on these results are presented. The requirement for an appropriate quality assessment methodology is also addressed.

MUIR, L.J., RICHARDSON, I.E.G., HAMILTON, K., 2005. Visual perception of content-prioritised sign language video quality. *Proceedings of the IEE international conference on Visual Information Engineering*, University of Glasgow, Glasgow. 4-6 April, pp.17-22

Notes: The IET Visual Information Engineering conferences bring together researchers, developers, creators, educators, and practitioners in image processing, machine vision, computer graphics, virtual and augmented environments, and visual communications to share their latest achievements and explore future directions and synergies in these fields. A paper was presented at the 2005 conference in Glasgow which presented a new method of subjective testing of perceived quality of selectively prioritised video for sign language communication.

Abstract: Video communication systems currently provide poor quality and performance for deaf people using sign language, particularly at low bit rates. Our previous work, involving eye movement tracking experiments and analysis of visual attention mechanisms for sign language, demonstrated a consistent characteristic response which could be exploited to enable optimisation of video coding systems performance by prioritising content for deaf users. This paper describes an experiment designed to test the perceived quality of selectively prioritised video for sign language communication. A series of selectively

degraded video clips was shown to individual deaf viewers. Participants subjectively rated the quality of the modified video on a Task-Specific Category Rating (TSCR) scale developed for sign language users. The results demonstrate the potential to develop content-prioritised coding schemes, based on viewing behaviour, which can reduce bandwidth requirements and provide best quality for the needs of the user. We propose selective quantisation to reduce compression in visually important regions of video images, which require spatial detail for small slow motion detection, and increased compression of regions regarded in peripheral vision where large rapid movements occur in sign language communication.

MUIR, L.J., RICHARDSON, I.E.G., LEAPER, S., 2003. Gaze tracking and its application to video coding for sign language. *Proceedings of the twenty-third International Picture Coding Symposium*, St. Malo, France. 22-26 April 2003, pp. 321-325

Notes: Picture Coding Symposium (PCS) is an international conference (sponsored by INRIA-Rennes in cooperation with the IEEE and EURASIP) devoted specifically to picture coding for industry, research, academia and users. A paper was presented at the 23rd international PCS which presented a novel method of video coding - content-prioritised coding for sign language video communication.

Abstract: Sign language communication via videotelephone has demanding visual quality requirements. In order to optimise video coding for sign language it is necessary to quantify the importance of areas of the video scene. Eye movements of deaf users are tracked whilst watching a sign language video sequence. The results indicate that the gaze tends to concentrate on the face region with occasional excursions (saccades). The implications of these results for prioritised coding of sign language video sequences are discussed.

Journal paper

MUIR, L. J. AND RICHARDSON, I. E. G., 2005. Perceptions of sign language and its application to visual communications for deaf people. *Journal of Deaf Studies and Deaf Education*, *10*(4), pp. 390-401

Notes: The Journal of Deaf Studies and Deaf Education is a peer-reviewed scholarly journal integrating and coordinating applied research relating to individuals who are deaf on issues of current and future concern to allied fields, encouraging interdisciplinary discussion. This paper presented novel eye movement tracking research in the context of developing new video communication systems for deaf people. The paper presented innovative and multidisciplinary research of interest to the engineering and social sciences research communities as well as the wider deaf community and its support organisations.

Abstract: Video communication systems for deaf people are limited in terms of quality and performance. Analysis of visual attention mechanisms for sign language enabled optimisation of video coding systems for deaf users. Eye movement tracking experiments were conducted with profoundly deaf volunteers while watching sign language video clips. Deaf people were found to fixate mostly on the facial region of the signer to pick up small detailed movements associated with facial expression and mouth shapes. Lower resolution, peripheral vision is used to process information from larger, rapid movements of the signer in the video clips. A coding scheme which gives priority to the face of the signer is applied to improve perception of video quality for sign language communication.

Contributions to Patents and Awards

Patent application: Variable bilateral filter-based content prioritised video coding.

Scottish Enterprise Proof of Concept Award 2006/7: Perceptually optimised video coding.

Appointments

Member of the Office of Communications (Ofcom) Advisory Committee on Older and Disabled People (Appointed on 1 April 2007).

Appendix B: ViewPoint™ EyeTracker Technical Specifications (Arrington Research, 2002)

Item	Specifications
Tracking Method	Infrared video. Bright pupil or dark pupil. Monocular.
Software	PC
Measurement principle	The user can select between three methods: Pupil only, corneal reflection only, or both together for greater tolerance to head movements.
Accuracy	Approximately 0.25° - 1.0° visual arc
Spatial resolution	Approximately 0.15° visual arc
Temporal resolution	Selectable by the user between 60 Hz and 30 Hz.
Allowable head movement	Small movements allowed. Subject's pupil and corneal reflection must remain within the camera image.
Visual range	Horizontal: +/- 44° of visual arc Vertical: +/- 20° of visual arc
Pupil size resolution	Measures pupil height and width to better than 0.03 mm instantaneous (no averaging).
Calibration	ViewPoint™ starts in a roughly calibrated state that is adequate for determining screen quadrants or other relative movement measurement such as objective preference-of-looking tasks. For accurate position of gaze, calibration is required once per subject. New subject setup time between 1-5 minutes. Calibration settings can be stored and reused each time a subject returns. Slip Correction feature and re-presentation of stray calibration points.
Auto threshold	The program scans over the video image for the pupil and / or for the corneal reflection. The luminance threshold for discriminating these can be adjusted. The auto threshold feature provides good threshold levels automatically. Little or no manual adjustment required.
Blink suppression	Automatic blink detection and suppression.
Data recorded	Eye data: X, Y position of gaze, pupil height and width, ocular torsion, delta time, total time, and regions of interest (ROI). Asynchronous records include: State transition markers, key presses and data from other programs. Data is stored in ASCII files.
Real-time communication	Same computer: Software Developers Kit (SDK) supplies everything required for seamless interface between ViewPoint™ and an external program. This includes: DLL with shared memory, .h and .lib files plus sample source code written in C Language. Serial port: Sends eye data packets and asynchronous packets equivalent to information in ASCII data files at rates of up to 56K. Receive real time data from other programs and store it asynchronously into data files. AnalogOut option: Selectable unipolar or bipolar voltage ranges: +/- 10, 5, 2.5 Selectable data items: position of gaze (x,y), pupil (h,w), velocity (dx,dy), and raw pupil, glint or vector data. TTL capabilities. TTL

	in/out option: Eight TTL input channels are interfaced to place marker codes into the ViewPoint™ data file. Eight TTL output channels that indicate when the position of gaze is inside ViewPoint™ region of interest areas ROI-0 to ROI-7.
Real-time display	Gaze point history, gaze trace, fixation duration, pupil size and ROIs, can be graphically displayed over stimulus image. Visible to the user and / or the subject. Real-time pen plots of X and Y position of gaze, velocity, ocular torsion, pupil width and pupil aspect ratio.
Stimulus Presentation	Pictures and movies can be displayed in full stimulus windows or in user specified ROIs. Auditory cues can be integrated. Gaze contingent stimulus presentation via state logic.
System requirements:	OS: Windows 2000, XP Machine: Pentium compatible
OEM Support	The ViewPoint™ software can be configured to accommodate a wide variety of OEM hardware. Resolution and accuracy will then depend on the OEM camera and hardware configuration.

Appendix C: Correlation-Based Optical Flow Estimation Algorithm (Matlab code)

```
%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%
% correl_flow.m
% Correlation-based optical flow estimation
% IGR & LJM, 23-06-05, reads 24-bit ("full colour") Bitmap files
%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%

filename = input('Bitmap 1 (24-bit):', 's');
B1 = imread(filename);
width = size(B1,2);
height = size(B1,1);
filename = input('Bitmap 2 (24-bit):', 's');
B2 = imread(filename);
if (size(B1) ~= size(B2))
   fprintf(1,'Error: frame sizes do not match\n');
   return;
end

figure;
subplot(1,2,1);imshow(B1);
subplot(1,2,2);imshow(B2);
B1a = double(B1);
B2a = double(B2);      % Convert to double precision
Y1 = zeros(height, width);      % Luminance frame 1
Y2 = zeros(height, width);      % Luminance frame 2
for i=1:height              % Convert RGB bitmap to luminance array
     for j=1:width
        Y1(i,j) = round((B1a(i,j,1)+B1a(i,j,2)+B1a(i,j,3))/3);
        Y2(i,j) = round((B2a(i,j,1)+B2a(i,j,2)+B2a(i,j,3))/3);
     end
end

% Area to compare: current position +/- (region) pixels (e.g. region = 2: compare 5x5 area)
region = 2; searchwin = 2; % +/- search area
xvec = zeros(height, width);    % Vectors in x-direction
yvec = zeros(height, width);    % Vectors in y-direction
saemap=zeros(height, width);
zeropref = 10;
for x = (1+searchwin+region):(width-searchwin-region)
   for y = (1+searchwin+region):(height-searchwin-region)
      minsae = 1000000;  % Minimum sae between blocks under comparison
      bestdx = 0; bestdy = 0;
      curblock = Y2(y-region:y+region, x-region:x+region);
      for dx = (-searchwin):searchwin
         for dy = (-searchwin):searchwin
            prevblock = Y1(y-region+dy:y+region+dy, x-region+dx:x+region+dx);
            cursae = sum(sum(abs(curblock-prevblock)));
```

```
            if (dx ~= 0) | (dy ~= 0)
               cursae = cursae + zeropref; % Prefer 0,0 vector
            end
            if (cursae < minsae)
               minsae = cursae;
               bestdx = dx;
               bestdy = dy;
            end
         end
      end
      xvec(y, x) = bestdx;
      yvec(y, x) = bestdy;
      saemap(y, x) = minsae;
   end
   fprintf(1, '.');
end

figure;
% Subsample and plot
subsample = 2;
xout = zeros(floor(height/subsample), floor(width/subsample));
yout = zeros(floor(height/subsample), floor(width/subsample));
for y = 1:floor(height/subsample)
   for x = 1:floor(width/subsample)
      xout(y,x) = xvec(y*subsample,x*subsample);
      yout(y,x) = yvec(y*subsample,x*subsample);
   end
end
quiver(yout, xout, 'k');
axis ij;
```

Appendix D: Phase-Based Optical Flow Estimation Algorithm (Matlab code)

```
%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%
% ofrun.m
% Computation of the Gautama and van Hulle (2002) phase-based optical flow
% for an image sequence including visualisation of subsampled flow fields.
% LJM, 23-06-05, reads CIF file, number of frames and temporal span (ts).
%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%

clear all

%%%%%%%%%%%%%%%%%%
% Set Parameters %
%%%%%%%%%%%%%%%%%%

gx = 30;          % Number of evaluation points in X dimension =25 for read sequence
thres_lin = 1;    % Linearity threshold for phase gradient 0.01
nc_min = 7;       % Minimum number of valid component velocities 7

%%%%%%%%%%%%
% Open CIF File%
%%%%%%%%%%%%

% get the CIF file
[fname, dirname] = uigetfile('*.*', 'Select CIF video file');
fnameanddir = sprintf('%s%s', dirname, fname);
infile = fnameanddir;
fprintf(1, 'File name = %s\n', infile);

%%%%%%%%%%%%%%%%%%
% Get Variables %
%%%%%%%%%%%%%%%%%%

% get the number of frames to process
a = inputdlg('Enter number of frames (0 to use max from file)')
nframes = str2num(a{1});
fprintf(1, 'Frames = %d\n', nframes);

mid1 = floor(nframes/2);
fprintf(1, 'Mid1 = %d\n', mid1);
if (rem(mid1,2)==0)
        mid = mid1+1;
else
        mid = mid1-1;
end
fprintf(1, 'Mid = %d\n', mid);

% get the temporal span (ts)
b = inputdlg('Enter the required temporal span')
```

```
ts = str2num(b{1});
fprintf(1, 'Temporal span, ts = %d \n', ts);

%%%%%%%%%%%%%%%%
% Read Sequence %
%%%%%%%%%%%%%%%%

II = read_cif(infile, nframes, mid, ts);
[sy sx ts] = size(II);

%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%
% Compute Optical Flow Field %
%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%

O = optical_flow (II, gx, thres_lin, nc_min);

%%%%%%%%%%%%%%%%%
% Plot %
%%%%%%%%%%%%%%%%%

mag = 1;
pcolor (II(:,:,floor(ts/2)+1));
shading flat
colormap gray
axis image
hold on
% Point where Gaussian envelope drops to 10%
offset = ceil(9.31648319*sqrt(log(100)));
[Vx Vy] = vis_flow (O(:,:,1), O(:,:,2), gx, offset, 1, 'm');
hold off

%%%%%%%%%%%
% Evaluate %
%%%%%%%%%%%

% Coverage
if (gx==0)
        jmp = 1;
else
        jmp = floor(sx/gx);
        jmp = jmp + (jmp==0);
end
nv = length(offset+1:jmp:(sy-offset))*length(offset+1:jmp:(sx-offset));
fprintf ('Coverage: %.2f %%\n', sum(sum(~isnan(Vx)))/nv*100);
```

```
%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%
% read_cif.m
% Read CIF image sequence into 3-D array (luminance only) for optical
% flow input
% LJM, 23-06-05.
%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%

function II = read_cif (infilename, nframes, mid, ts, sx, sy)

width = 352; height = 288;  % CIF width and height of each frame in pixels

fid = fopen(infilename, 'r');
fprintf(1, 'FID: %d\n', fid);
d = dir(infilename);
filebytes = d.bytes;
numframes = filebytes / (width * height * 1.5); % Assume 4:2:0 sampling

if (nframes == 0)
   nframes = numframes;
end
fprintf(1, 'File %s, loading %d frames\n', infilename, nframes);

%II = zeros(height, width, nframes);    % y, x, time for optical flow algo
II = zeros(height, width, ts);    % y, x, time for optical flow algo

aux1 = floor(ts/2);

if (rem(ts,2)==0)
        IND = mid-aux1-1:mid+aux1;
else
        IND = mid-aux1:mid+aux1;
end

fprintf(1, 'IND = %d \n', IND);

skip = mid-(aux1+1);
fprintf(1, 'skip = %d \n', skip);

fread(fid, width*height*1.5*skip, 'uchar');
%for i=1:nframes
for i=1:ts
   for j = height:-1:1    % for each line : Y (luminance)
      II(j, :, i) = rot90(fread(fid, width, 'uchar'));
   end
   fread(fid, width*height*0.5, 'uchar');   % Skip U, V data in file

end
fclose(fid);
```

```
%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%
% optical_flow.m
% Original Gautama and van Hulle (2002) Phase-based Optical Flow Algorithm
%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%

%       Parameters:
%       II [sy sx ts]       Image Sequence (Y-X-t)
%       gx                  Number of velocity vectors along X-axis (0=all)
%       thres_lin           Linearity threshold [.05]
%       nc_min              Minimal number of valid component velocities for
%                           computation of full velocity [5]

function O = optical_flow (II, gx, thres_lin, nc_min)

if (nargin<1)
        error ('Please provide an input sequence');
end
if (nargin<2)
        gx = 0;
end
if (nargin<3)
        thres_lin = .05;
end
if (nargin<4)
        nc_min = 5;
end
[sy sx ts] = size(II);

if (gx==0)
        jmp = 1;
else
        jmp = floor(sx/gx);
        jmp = jmp + (jmp==0);
end

%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%
% Load Filterbank Parameters %
%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%

W = [   0.02156825 -0.08049382; ...
        0.05892557 -0.05892557; ...
        0.08049382 -0.02156825; ...
        0.08049382 0.02156825; ...
        0.05892557 0.05892557; ...
        0.02156825 0.08049382; ...
        0.06315486 -0.10938742; ...
        0.10938742 -0.06315486; ...
        0.12630971 0.00000000; ...
        0.10938742 0.06315486; ...
        0.06315486 0.10938742];
```

```
S = [9.31648319 9.31648319 9.31648319 9.31648319 9.31648319 9.31648319 ...
        6.14658664 6.14658664 6.14658664 6.14658664 6.14658664]';
nn = size(W,1);

%%%%%%%
% Aux %
%%%%%%%

xx = (1:ts);
xx3 = zeros(1,1,ts);
xx3(1:ts) = 1:ts;
Sxx = sum(xx.^2);
Sx = sum(xx);
den = (ts.*Sxx-Sx.^2);
pi2 = 2*pi;

%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%
% Compute Filter Outputs & Component Velocities
%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%

tclock1 = clock;
AC = zeros(sy,sx,ts);
AS = zeros(sy,sx,ts);
FV = zeros(nn,sy,sx); % Filter Component Velocity
LE = zeros(nn,sy,sx); % MSE of Regression
Ec = zeros(sy,sx,nn);
offset = ceil(max(S)*sqrt(log(100)));   % Point where Gaussian envelope drops to 10%
%offset=0;
[offx offy] = meshgrid(1:sx,1:sy);
for n=1:nn,fprintf('+');end;
for i=offset+1:jmp:(sy-offset),fprintf('.');end;fprintf('\n');

for n=1:nn
        % Generate 1D kernels
        win_len = floor(6*S(n));
        cx = (0:win_len-1)'-win_len/2+.5;
        cx2 = pi2.*cx;
        G = exp(-cx.^2./(2*S(n)*S(n)))./(sqrt(2*pi)*S(n));
        FCx = G.*cos(W(n,1).*cx2);
        FCy = G.*cos(W(n,2).*cx2);
        FSx = G.*sin(W(n,1).*cx2);
        FSy = G.*sin(W(n,2).*cx2);

        % Exceptions for null frequencies
        if (sum(FCx.^2)==0)
                FCx = ones(win_len,1);
        end
        if (sum(FCy.^2)==0)
                FCy = ones(win_len,1);
        end
```

```
        if (sum(FSx.^2)==0)
                FSx = ones(win_len,1);
        end
        if (sum(FSy.^2)==0)
                FSy = ones(win_len,1);
        end

        % Perform Convolutions, room for improvement (subsampling)
        for t=1:ts
                Ilp = II(:,:,t)';

                % Sine Filter
                Tsx = conv2(Ilp, FSx, 'same')';
                T2 = conv2(Tsx, FCy, 'same');
                Tcx = conv2(Ilp, FCx, 'same')';
                T4 = conv2(Tcx, FSy, 'same');
                AS(:,:,t) = T2 + T4;

                % Cosine Filter
                %Tcx = conv2(Ilp, FCx, 'same')';
                T2 = conv2(Tcx, FCy, 'same');
                %Tsx = conv2(Ilp, FSx, 'same')';
                T4 = conv2(Tsx, FSy, 'same');
                AC(:,:,t) = T2 - T4;
        end

        % Compute and Unwrap Phase
        Mcos = (AC==0);
        P = atan(AS./(AC+Mcos))+pi.*(AC<0);
        P(Mcos) = NaN;
        k = 2;
        while (k<=ts)
                D = P(:,:,k) - P(:,:,k-1);
                A = abs(D)>pi;
                P(:,:,k:ts) = P(:,:,k:ts) - repmat(pi2.*sign(D).*A,[1 1 ts-k+1]);
                k = k + (sum(sum(A))==0);
        end

        % Compute Filter Component Velocity
        Sxy = sum(repmat(xx3,[sy sx 1]).*P,3);
        Sy = sum(P,3);
        a = (Sxx.*Sy-Sx.*Sxy)./den;
        b = (ts.*Sxy-Sx.*Sy)./den;
        Reg = repmat(a,[1 1 ts])+repmat(b,[1 1 ts]).*repmat(xx3,[sy sx 1]);
        LE(n,:,:) = mean((Reg-P).^2,3)./abs(b+(b==0));
        FV(n,:,:) = -b./(pi2*sum(W(n,:).^2)).*(W(n,1)+sqrt(-1)*W(n,2));

        fprintf ('*');

end
```

```
tclock2 = clock;
time1 = etime(tclock2,tclock1);

%%%%%%%%%%%%%%%%%%%%%%%%%
% Compute Full Velocity %
%%%%%%%%%%%%%%%%%%%%%%%%%

O = repmat(NaN, [sy sx 2]);
for i=offset+1:jmp:(sy-offset)
  for j=offset+1:jmp:(sx-offset)
        % Linearity Check
        IND1 = find(LE(:,i,j)<thres_lin);
        V = FV(IND1,i,j);
        nc = length(IND1);
        if (nc>=nc_min)
              L_2 = V.*conj(V);
              X = real(V);
              Y = imag(V);
              sumX = sum(X);
              sumY = sum(Y);
              sumXYL_2 = sum(X.*Y./L_2);
              sumXXL_2 = sum(X.^2./L_2);
              sumYYL_2 = sum(Y.^2./L_2);
              den = (sumXYL_2^2-sumXXL_2*sumYYL_2);
              xr = -(sumX*sumYYL_2-sumY*sumXYL_2) / den;
              yr = (sumX*sumXYL_2-sumY*sumXXL_2) / den;
              O(i,j,:) = [xr yr];
        end
  end
  fprintf ('*');
end
fprintf ('\n');
tclock3 = clock;
time2 = etime(tclock3, tclock2);
fprintf ('\tElapsed time: %.2f + %.2f = %.2f [sec]\n', time1, time2, time1+time2);

%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%
% vis_flow.m
% Original quiver, with subsampling
%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%

function [Ox,Oy] = vis_flow (VVx, VVy, gx, offset, mag, col);

if (nargin<3)
        gx = 25;
end
if (nargin<4)
        offset = 0;
end
if (nargin<5)
```

```
        mag = 1;
end
if (nargin<6)
        col = 'b';
end

[sy sx] = size(VVx);
if (gx==0)
        jmp = 1;
else
        jmp = floor(sx/gx);
        jmp = jmp + (jmp==0);
end

indx = (offset+1):jmp:sx;
c = 1;
CX = [];
CY = [];
for j=(1+offset):jmp:sy
        Vx(c,:) = VVx(j,indx);
        Vy(c,:) = VVy(j,indx);
        CX(c,:) = indx;
        %CY(c,:) = ones(size(indx)).*(sy-j+1);
        CY(c,:) = ones(size(indx)).*j;
        c = c+1;
end

if (isnan(Vx(1,1)))
        Vx(1,1) = 1;
        Vy(1,1) = 0;
        CX(1,1) = 1;
        CY(1,1) = 1;
end

M = ~isnan(Vx) & ~isnan(Vy);
H = quiver (CX(M), CY(M), Vx(M), Vy(M), mag);
s = size(VVx);
axis ([0 s(2) 0 s(1)]);
set (H, 'Color', col);

switch nargout
        case 0
                clear Ox;
                clear Oy;
        case 1
                Ox = H;
        otherwise
                Ox = Vx;
                Oy = Vy;
end
```

Appendix E: Spatial Video Image Foveation Algorithm (Matlab code)

```
%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%
% procfoveate.m
% IGR/LJM 17-02-2004: Processes avi sequence and foveates video image
% data using x,y coordinates (obtained using getsignxy.m) as the centre point
% for foveation.
%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%

% Get AVI file
[fname, dirname] = uigetfile('*.avi', 'Select input AVI file');
fnameanddir = sprintf('%s%s', dirname, fname);
vidfname = fnameanddir;
fileinfo = aviinfo(vidfname);
if fileinfo.VideoCompression ~= 'none'
   errordlg(sprintf('Error: %s uses unsupported videocompression %s', fnameanddir,
fileinfo.VideoCompression));
   return;
end

% Get foveation centre (x,y coordinates obtained from getsignxy.m)
fname = uigetfile('*.mat', 'Select MAT file');  % should contain one (x,y) pair for each
frame in file
load(fname);
[fname, dirname] = uiputfile('*.avi', 'Enter output AVI file');
fnameanddir = sprintf('%s%s', dirname, fname);
aviobj = avifile(fnameanddir, 'COMPRESSION', 'None');
aviobj.fps = fileinfo.FramesPerSecond;

% Constants and Variables
levels = 7;                         % number of pyramid levels
CT0 = 1/40;                         % variable minimum contrast threshold
alpha = 0.106;                      % spatial frequency decay constant
epsilon2 = 2.3;                     % half resolution eccentricity constant
dotpitch = .25*(10^-3);             % monitor dot pitch in meters
viewingdist = 0.305;                % variable viewing distance in meters

for framenum = 1:length(xarray)
   % read a frame
   mov = aviread(vidfname, framenum);
   inimage = double(mov.cdata);  % colour image, RGB
   % set x, y of fovea centre
   fovx = xarray(framenum);
   fovy = yarray(framenum);
   outimage = avifoveate(inimage, fovx, fovy, levels, CT0, alpha, epsilon2, dotpitch,
viewingdist);
   outframe.cdata = uint8(outimage * 255);
   outframe.colormap = [ ];
   aviobj = addframe(aviobj, outframe);
   fprintf(1, '.');
```

```
   if mod(framenum, 20) == 0
      fprintf(1, '\n');
   end
end
fprintf(1, '\n');
aviobj = close(aviobj);
msgbox(sprintf('Completed writing file %s', fnameanddir));

%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%
% avifoveate.m
% Creates foveated video images based on Gaussian multiresolution
% pyramid using original Geisler and Perry (1998) formula.
%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%

function outimage = avifoveate(inimage, fovx, fovy, levels, CT0, alpha, epsilon2,
dotpitch, viewingdist);

% normalize values to maximum value; for most MATLAB
% functions, this value must be one.
max_cval = 1.0;
%range = [min(inimage(:)) max(inimage(:))];
range = [0 255];
img_size = size(inimage(:,:,1));
inimage = max_cval.*(inimage-range(1)) ./ (range(2)-range(1));

% ex and ey are the x- and y- offsets of each pixel compared to
% the point of focus (fovx,fovy) in pixels.
[ex, ey] = meshgrid(-fovx+1:img_size(2)-fovx,-fovy+1:img_size(1)-fovy);

% eradius is the radial distance between each point and the point
% of gaze.  This is in meters.
eradius = dotpitch .* sqrt(ex.^2+ey.^2);
clear ex ey;

% calculate ec, the eccentricity from the foveal centre, for each
% point in the image.  ec is in degrees.
ec = 180*atan(eradius ./ viewingdist)/pi;

% maximum spatial frequency (cpd) which can be accurately represented onscreen:
maxfreq = pi ./ ((atan((eradius+dotpitch)./viewingdist) - ...
            atan((eradius-dotpitch)./viewingdist)).*180);

clear eradius;

% calculate the appropriate (fractional) level of the pyramid to use with
% each pixel in the foveated image.

% eyefreq is a matrix of the maximum spatial resolutions (in cpd)
% which can be resolved by the eye
eyefreq = ((epsilon2 ./(alpha*(ec+epsilon2))).*log(1/CT0));
```

```
% pyrlevel is the fractional level of the pyramid which must be
% used at each pixel in order to match the foveal resolution
% function defined above.
pyrlevel = maxfreq ./ eyefreq;

% constrain pyrlevel in order to conform to the levels of the
% pyramid which have been computed.
pyrlevel = max(1,min(levels,pyrlevel));

clear ec maxfreq;

% Plot the foveation region matrix (optional)
%figure(2);
%pcolor(levels - pyrlevel);axis ij; colormap gray; shading flat;
%title('Foveation Region: white means higher resolution');

% create storage for the final foveated image
outimage = zeros(img_size(1),img_size(2),3);

% create matrices of x&y pixel values for use with interp3
[xi,yi] = meshgrid(1:img_size(2),1:img_size(1));

% Repeat foveation procedure 3 times; once for each
% of the three colour planes (in the optimised version for the Proof of Concept project
% this code is only executed once for the luminance data.)
for color_idx = 1:3
  img = inimage(:,:,color_idx);

  % build Gaussian pyramid
  [pyr,indices] = buildGpyr(img,levels);

  % upsample each level of the pyramid in order to create a foveated representation
  point = 1;
  blurtree = zeros(img_size(1),img_size(2),levels);
  for n=1:levels
    nextpoint = point + prod(indices(n,:)) - 1;
    show = reshape(pyr(point:nextpoint),indices(n,:));
    point = nextpoint + 1;
    blurtree(:,:,n) = ...
          imcrop(upBlur(show, n-1),[1 1 img_size(2)-1 img_size(1)-1]);
  end

  clear pyr indices;
  clear show;

  % create foveated image by interpolation
  outimage(:,:,color_idx) = interp3(blurtree,xi,yi,pyrlevel, '*linear');

end
```

```
clear inimage img;
clear xi yi;

%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%
% getsignxy.m
% IGR 1.04, Extracts the x,y coordinates for the foveation centre.
%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%

[fname, dirname] = uigetfile('*.avi', 'Select AVI file');
fnameanddir = sprintf('%s%s', dirname, fname);
vidfname = fnameanddir;
fileinfo = aviinfo(fnameanddir);
if fileinfo.VideoCompression ~= 'none'
   errordlg(sprintf('Error: %s uses unsupported videocompression %s', fnameanddir,
fileinfo.VideoCompression));
   return;
end
framenum = 0;
xarray = [ ]; yarray = [ ]; x= 0; y = 0;
a = inputdlg('Enter number of frames (0 to use max from file)')
nframes = str2num(a{1});
if (nframes == 0)
   nframes = fileinfo.NumFrames;
end
while (framenum < nframes)
   mov = aviread(vidfname, framenum+1);
   h = image(mov.cdata);
   hold on;
   if (framenum > 0)
      plot(x, y, 'yx');
   end
   hold off;
   title(sprintf('Frame %d / %d', framenum, nframes));
   [x,y,button] = ginput(1);
   x = round(x); y = round(y);
   xarray(framenum+1) = x; yarray(framenum+1) = y;
   fprintf(1, '%d, %d\n', x, y);
   if (button == 1)
      framenum = framenum + 1;
   else
      framenum = framenum - 1;
   end
   if (framenum < 0)
      framenum = 0;
   end
end
button = questdlg('Click to play back', '', 'OK', 'OK');
for i=1:nframes
   mov = aviread(vidfname, i);
   h = image(mov.cdata);
```

```
    hold on;
    plot(xarray(i), yarray(i), 'yx');
    title(sprintf('Frame %d / %d', i, nframes));
    hold off;
    pause(0.1);
end

save xy xarray yarray
msgbox('X, Y arrays saved in xy.mat'); % Rename xy.mat for each video sequence.
```

Appendix F: Foveation-Weighted Temporal Filter Algorithm (Matlab code)

```
%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%
% batchtempfilt.m
% Foveation-Weighted Temporal Filter (FWTF)
% LJM, 10-11-05, batch file for FWTF of input avi at a range of Filter Strengths (FS)
%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%

% Useage: procfwtf (input filename, output filename, x,y data file, FS)
procfwtf ('LisaTVOut.avi', 'LisaTVfs2_gaus.avi', 'LisaTVxy.mat',  2);
procfwtf ('LisaTVOut.avi', 'LisaTVfs4_gaus.avi', 'LisaTVxy.mat',  4);
procfwtf ('LisaTVOut.avi', 'LisaTVfs6_gaus.avi', 'LisaTVxy.mat',  6);
procfwtf ('LisaTVOut.avi', 'LisaTVfs8_gaus.avi', 'LisaTVxy.mat',  8);
procfwtf ('LisaTVOut.avi', 'LisaTVfs10_gaus.avi', 'LisaTVxy.mat',  10);

%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%
% procfwtf.m
% Processes avi sequence and applies FWTF to video image data using x,y
% coordinates (obtained using getsignxy.m) as the centre point for foveation.
% LJM 07-09-2005
%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%

function procfwtf (inavi, outavi, inxy, filtstrength);

vidfname = inavi;
fileinfo = aviinfo(vidfname);
if fileinfo.VideoCompression ~= 'none'
   errordlg(sprintf('Error: %s uses unsupported videocompression %s', fnameanddir,
fileinfo.VideoCompression));
   return;
end
fname = inxy;  % should contain one (x,y) pair for each frame in file
load(fname);
aviobj = avifile(outavi, 'COMPRESSION', 'None');
aviobj.fps = fileinfo.FramesPerSecond;
FS = filtstrength;

% Geisler & Perry (1998) Constants and Variables
levels = 7;                         % number of pyramid levels
CT0 =  1/75;                        % minimum contrast threshold
alpha = 0.106;                      % spatial frequency decay constant
epsilon2 = 2.3;                     % half resolution eccentricity constant
dotpitch = .25*(10^-3);             % monitor dot pitch in meters
viewingdist = 0.305;                % viewing distance in meters

% FWTF Variables
filtkernel = [1 0 0 0]; % Default filtkernel (no filter)

% set number of frames to read
nframes = length(xarray); % read all frames
```

```
% nframes = 262;  % optional code to read set number of frames
fprintf(1, 'Reading %d frames', nframes);

% read data for all frames into a frame data store
framestore = zeros(nframes, fileinfo.Height, fileinfo.Width, 3);
tic  % start timer for processing all frames
for frame = 1:nframes
   % read a frame
   framedata = aviread(vidfname, frame);
   inframe = double(framedata.cdata);  % colour image, RGB
   % store it for ref to previous frames in the temp filter process
   framestore(frame, :, :, :) = double(inframe); % orig %framestore(frame, :, :, :) =
double(inframe * 255);
   fprintf(1, '.');
end
fprintf(1, '\nFinished Reading %d frames\n', nframes);

% process each frame in the sequence
for framenum = 1:nframes
   % set x, y of fovea centre
   fovx = xarray(framenum);
   fovy = yarray(framenum);

   % call FWTF
   outimage = fwtf(FS, filtkernel, framestore, framenum, fovx, fovy, levels, CT0, alpha,
epsilon2, dotpitch, viewingdist);

   %send data to ouput avi file
   outframe.cdata = uint8(outimage);
   outframe.colormap = [ ];
   aviobj = addframe(aviobj, outframe);
   fprintf(1, '.');
   if mod(framenum, 20) == 0
      fprintf(1, '\n');
   end
end
fprintf(1, '\n');
toc % stop timer for processing all frames
t=toc;
fprintf(1, '\nt=%d \n', t);
aviobj = close(aviobj);
fprintf(1,'Completed writing file %s', outavi);
```

```
%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%
% fwtf.m
% Creates FWTF video images based on Gaussian multiresolution
% pyramid using original Geisler and Perry (1998) formula and FS parameter.
% LJM 07-09-2005
%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%

function outimage = fwtf(FS, filtkernel, framestore, framenum, fovx, fovy, levels, CT0, alpha, epsilon2, dotpitch, viewingdist);

% normalize values to maximum value; for most MATLAB
% functions, this value must be one.
max_cval = 1.0;
%range = [min(inimage(:)) max(inimage(:))];
range = [0 255];

% Filter constants and calculations
width = 352; height = 288;
nfilt = length(filtkernel); % Number of frames to filter
sfilt = sum(filtkernel);
inimage = zeros(1, height, width, 3);
img_size = size(inimage(:,:,:,1)); % reads a frame from framestore
inimage = max_cval.*(inimage-range(1)) ./ (range(2)-range(1));

% ex and ey are the x- and y- offsets of each pixel compared to
% the point of focus (fovx,fovy) in pixels.
[ex, ey] = meshgrid(-fovx+1:img_size(3)-fovx,-fovy+1:img_size(2)-fovy);
% eradius is the radial distance between each point and the point
% of gaze in meters.
eradius = dotpitch .* sqrt(ex.^2+ey.^2);
clear ex ey;

% calculate ec, the eccentricity from the foveal center, for each
% point in the image.  ec is in degrees.
ec = 180*atan(eradius ./ viewingdist)/pi;

% maximum spatial frequency (cpd) which can be accurately represented onscreen:
maxfreq = pi ./ ((atan((eradius+dotpitch)./viewingdist) - ...
          atan((eradius-dotpitch)./viewingdist)).*180);
clear eradius;

% calculate the appropriate (fractional) level of the pyramid to use with
% each pixel in the foveated image.

% eyefreq is a matrix of the maximum spatial resolutions (in cpd)
% which can be resolved by the eye
eyefreq = ((epsilon2 ./(alpha*(ec+epsilon2))).*log(1/CT0));

% pyrlevel is the fractional level of the pyramid which must be
% used at each pixel in order to match the foveal resolution
```

```
% function defined above.
pyrlevel = maxfreq ./ eyefreq;

% constrain pyrlevel in order to conform to the levels of the
% pyramid which have been computed.
pyrlevel = max(1,min(levels,pyrlevel));
clear ec maxfreq;

% create storage for the final FWTF image
outimage = zeros(img_size(2),img_size(3),3);

% Repeat FWTF procedure 3 times; once for each
% of the three colour planes.
for color_idx = 1:3
  % FWTF - Process one complete "component", e.g. Y or R/G/B
    for i = 1:height
        for j=1:width
            outsample=0;
            sigma=0;
            % Note: frames 1:nfilt-1 are not filtered
            if ((framenum < nfilt)|(pyrlevel(i,j)==1))
                outsample = framestore(framenum, i, j, color_idx);
            else
                % Obtain the Gaussian filter kernel based on the values
                % of sigma obtained from the pyrlevel and filter strength (FS) multiplier
                % see filtercalcs spreadsheet (Table 9.1) for values
                sigma = (((pyrlevel(i,j))-1) + (FS/10));
                filt = fspecial('gaussian', [1 7], sigma);
                filtkernel = filt(4:7);
                for f = 1:nfilt
                    outsample = outsample + (framestore( framenum - f + 1, i, j, color_idx) *
filtkernel(f));
                end
                %Normalise intensity
                sfilt = sum(filtkernel);
                outsample = round(outsample / sfilt);
            end
            outimage(i, j, color_idx) = outsample;
        end
    end % End of "component"
end
clear inimage img; % FWTF - End of processing: result is in outimage
```

CD-ROM Appendices

Appendix I: EMT Data for Subject 1 plotted on BSL Test Video Sequence 1

Appendix II: EMT Data Timelines

Appendix III: Encoded Bit Count Analysis of BSL Test Video Sequences