DISTRIBUTED MULTIMEDIA QUALITY : THE USER PERSPECTIVE

STEPHEN RICHARD GULLIVER

DISTRIBUTED MULTIMEDIA QUALITY: THE USER PERSPECTIVE

A thesis submitted for the degree of Doctor of Philosophy

by

Stephen Richard Gulliver

Department of Information Systems and Computing,
Brunel University.

August 2004

ABSTRACT

Distributed multimedia supports a symbiotic *infotainment* duality, i.e. the ability to transfer information to the user, yet also provide the user with a level of satisfaction. As multimedia is ultimately produced for the education and / or enjoyment of viewers, the user's-perspective concerning the presentation quality is surely of equal importance as objective Quality of Service (QoS) technical parameters, to defining distributed multimedia quality. In order to extensively measure the user-perspective of multimedia video quality, we introduce an extended model of distributed multimedia quality that segregates quality into three discrete levels: the *network-level*, the *media-level* and *content-level*, using two distinct quality perspectives: the user-perspective and the technical-perspective.

Since experimental questionnaires do not provide continuous monitoring of user attention, eye tracking was used in our study in order to provide a better understanding of the role that the human element plays in the reception, analysis and synthesis of multimedia data. Results showed that video content adaptation, results in disparity in user video eye-paths when: i) no single / obvious point of focus exists; or ii) when the point of attention changes dramatically.

Accordingly, appropriate technical- and user-perspective parameter adaptation is implemented, for all quality abstractions of our model, i.e. network-level (via simulated delay and jitter), media-level (via a technical- and user-perspective manipulated region-of-interest attentive display) and content-level (via display-type and video clip-type). Our work has shown that user perception of distributed multimedia quality cannot be achieved by means of purely technical-perspective QoS parameter adaptation.

ACKNOWLEDGEMENTS

I would like to especially thank my wife (Claire) and my supervisor (Dr. George Ghinea), for their continued help and support. In addition, I would also like to thank all of the students, family members, and friends who helped by participating in my perceptual experiments. God Bless!

DECLARATION

The following papers have been published (or have been submitted for publication) as a direct or indirect result of the research discussed in this thesis:

JOURNALS

1. Gulliver, S. R., and Ghinea, G., (2004). Eye-gaze and Content-dependent Region-of-Interest based Video Adaptation: A Perceptual Comparison. Submitted to *IEEE Transactions on Circuit and Systems for Video Technology*.

2. Gulliver, S. R., and Ghinea, G., (2004). Starts in their Eyes: What Eye-Tracking Reveal about Multimedia Perceptual Quality. *IEEE Transaction on System, Man and Cybernetics, Part A, Vol. 34* (4), pp. 472-482.

3. Gulliver, S. R., Serif, T., and Ghinea, G., (2004). Pervasive and Standalone Computing: the Perceptual Effects of Variable Multimedia Quality. *International Journal of Human Computer Studies, Vol. 60,* pp. 640-665.

REVIEWED CONFERENCES

1. Gulliver, S. R., and Ghinea, G., (2004). Region of Interest Displays: Addressing a Perceptual Problem? Accepted for *IEEE Sixth International Symposium on Multimedia Software Engineering.*

2. Gulliver, S. R., and Ghinea, G., (2004). An Eye Opener: Low Frame Rates do Not Affect Fixations. *IEEE International Conference on Multimedia and Expo*, Taipei, Taiwan.

3. Gulliver, S. R., and Ghinea, G., (2004). Changing Frame Rate, Changing Satisfaction. *IEEE International Conference on Multimedia and Expo*, Taipei, Taiwan.

4. Serif, T., Gulliver, S. R., and Ghinea, G., (2004) Perceptual Impact of Multimedia QoS: a Comprehensive Study of Pervasive and Traditional Computing Devices, *ACM Symposium on Applied Computing*, Nicosia, Cyprus, pp. 1580-1585.

LIST OF FIGURES

LIST OF TABLES

Table **Page**

TABLE OF CONTENTS

Page

CHAPTER 1

CHAPTER 2

CHAPTER 4

CHAPTER 5

CHAPTER 6

CHAPTER 1

Introduction

1.1. Distributed Multimedia Quality

Distributed multimedia relies on both the multi-sensory nature of humans and the ability of computers to store, manipulate and communicate primarily video and audio data. Distributed multimedia covers a range of applications, which reflects the symbiotic *infotainment* duality of multimedia, i.e. the ability to transfer information to the user, yet also provide the user with a level of satisfaction. As multimedia applications are ultimately produced for the education and / or enjoyment of human viewers, the user's-perspective concerning the presentation quality is surely of equal importance to defining distributed multimedia quality as objective Quality of Service (QoS) technical parameters. Accordingly, distributed multimedia quality, in our perspective, is deemed as having two main facets: *of perception* and *of service.* The former facet (QoP - Quality of Perception) considers the user-perspective, measuring the infotainment aspect of the presentation. The latter facet (QoS – Quality of Service) characterises the technical-perspective and represents the performance properties provided by the multimedia technology.

1.2. Measuring Distributed Multimedia Quality

Distributed multimedia quality is not defined by a "single monotone dimension" [VIR95], yet has traditionally been judged using numerous factors, which have been shown to influence user criteria concerning presentation excellence, e.g. delay or loss of frames, audio clarity, lip synchronisation during speech, as well as the general relationship between visual auditory components [APT95]. Although previous work considers aspects that influence distributed multimedia quality, presenting a truly extensive examination of the distributed multimedia multidimensional quality paradigm is complex. To realize this view in our study, we have developed a model of quality first introduced by Wikstrand [WIK03], which segregates quality into three discrete levels: the *network-*, the *media-* and the *content-levels.*

- The network-level is concerned with how data is communicated over the network and includes variation and measurement of parameters including: bandwidth, delay, jitter and loss.
- The media-level is concerned with how the media is coded for the transport of information over the network and / or whether the user perceives the video as being of good or bad quality. Media-level parameters include: frame rate, bit rate, screen resolution, colour depth and compression techniques.
- The content-level is concerned with the transfer of information and level of satisfaction between the video media and the user, i.e. level of enjoyment, ability to perform a defined task, or the user's assimilate critical information from a multimedia presentation.

At each quality abstraction defined in our model, quality parameters can be experimentally adapted, e.g. jitter at the network-level, frame rate at the media-level and finally display-type at the content-level. Similarly, at each level of the model, quality can be measured, e.g. percentage of loss at the network-level, user mean opinion score (MOS) at the media-level, and task performance at the content-level. To further differentiate studies, in line with the infotainment duality of multimedia, we incorporated two distinct quality perspectives in our work: the user-perspective and the technical-perspective.

- ***User-Perspective*:** The user-perspective can be adapted and measured at the media- and content-levels. The network-level does not facilitate the user-perspective, since user perception can not be measured at this low level abstraction.
- ***Technical-Perspective*:** Technical parameters can be adapted and measured at all quality abstractions.

In our work, we consider previous studies, which involve quality variation and measurement at the three levels of quality abstraction identified. Special attention was given to differentiate the two distinct quality perspectives (the technical- or user-perspective). In summary:

- **Network-Level:** Technical-perspective network-level variation of bit error, segment loss, segment order [GHI00], delay and jitter [CLA99; GHI00; PRO99] have been used to simulate QoS deterioration. Technical-perspective network-level measurement of loss [GHI00; KOO98], delay and jitter [WAN01], as well as allocated bandwidth [WAN01] have all been used to measure network-level quality performance.

- **Media-Level:** Technical-perspective media-level variation of video and audio frame rate [APT95; GHI98; KAW95; KIE97; MAS01; WIJ99; WIL00a; WIL00b], captions [GULL03], animation method [WIK02], inter-stream audio-video quality [HOL97], image resolution [KIE97], media stream skews [STE96; WIJ99], synchronisation [STE96] and video compression codecs [MAS01; WIN01] have been used to vary quality definition. User-perspective media-level variation requires user data feedback and is limited to attentive displays, which manipulates video quality around a user's point of gaze. Technical-perspective media-level measurement is based on linear and visual quality models [ARD94; LIN96; QUA02; TE094; VAN96; VAV96, VEH96; WAT98; WIN01 ; XIA00], with the exception of [WAN01] who uses output frame rate as the quality criterion. User-perspective media-level measurement of quality has been used when measuring user 'watchability' (receptivity) [APT95], user rating of video quality [GHI98; WIK02], comparison between streamed video against the non-degraded original video [PRO99; WIN01], continuous quality assessment [WIL00a; WIL00b] and participant annoyance of synchronisation skews [STE96].

- **Content-Level:** Technical-perspective content-level variation was used to vary the content of experimental material [GHI98; GUL03; MAS01; PRO99; RIM99; STE96] as well as the presentation language [STE96]. User-perspective content-level variation was used to measure the impact of user demographics [GUL03], as well as volume and type of microphone [WAS00] on overall perception of multimedia quality. Technical-perspective content-level measurement has to date, only included stress analysis [WIL00a; WIL00b]. User-perspective content-level measurement has measured 'watchability' (receptivity) [APT95], 'ease of understanding', 'recall', 'level of interest', 'level of comprehension' [PRO99], information assimilation [GHI98; GUL03], predicted level of information assimilation [GUL03] and enjoyment [GUL03; WIK02].

1.3. Problem Statement and Research Aim

Current multimedia communication systems are inherently unsuited for the transport of loss tolerant and delay intolerant multimedia data. Consequently objective multimedia quality, determined by resourse allocation, varies widely between users, making the individual the ultimate judge of multimedia quality - it is his/her perception of quality that ultimately determines what defines acceptable multimedia quality. It is therefore only by considering the human element in multimedia communications that a true end-to-end quality-assured architectures will be possible. Moreover, previous research has indicated that if perceptual quality requirements are taken into account in the multimedia transmission, then potential resource savings can be obtained, resulting in more efficient and streamlined communication mechanisms that will take into account user requirements, the nature of the data being transported, as well as the networking environment over which this communication takes place.

To extensively measure the user's perception of multimedia quality, user perception should be measured across a range of quality abstractions. Variation of relevant technical- and user-perspective parameters is required at all quality abstractions: network-level (technical-perspective), media-level (both technical- and user-perspectives), and content-level (both technical- and user-perspectives). In addition, user perception of multimedia quality should consistently consider both how the multimedia presentation was assimilated / understood by the user at the content-level, yet also examine the user's satisfaction (both his/her satisfaction with the objective QoS setting and level of enjoyment concerning the video content). Interestingly, none of the mentioned studies achieved this set of criteria and it is on this problem that our research shall focus its attention.

In our work, we define the following research aim: to extensively consider the user's perception of distributed multimedia quality, by adapting relevant technical- and user-perspective parameters at all quality abstractions: network-level (technical-perspective), media-level (both technical- and user-perspectives), and content-level (both technical- and user-perspectives). In addition, user perception of multimedia quality should consistently consider both how the multimedia presentation was assimilated / understood by the user at the content-level, yet also examine the user's satisfaction (both his/her satisfaction with the objective QoS settings {media-level} and level of enjoyment concerning the video content {content-level}). To achieve the defined research aim, a series of three investigations, structured along the Network, Media and Content levels of our model, will be carried out, each targeting a major research objective of our work. To ensure consistent perceptual measurement, an adapted version of Quality of Perception (QoP) [GHP00] will be used in all investigations.

- **Objective 1: Measurement of the perceptual impact of network-level parameter variation.** To this end, we intend to measure the impact of delay and jitter on user perception of multimedia quality. In addition to QoP, eye tracking will be employed in our work. Eye-tracking systems are used as either a data-gathering device or can provide the user with interactive functionality [ISO00; REI02]. Depending on the equipment, eye-tracking devices can be considered as either intrusive or non-intrusive in nature [GOL02], can be developed as either a pervasive [SOD02] or standalone systems, and may have a level of immersion, which is perceived as being either high [HAY02] or low [PAS00]. Interpretation of eye movement data is based on the empirically validated assumption that when a person is performing a cognitive task, the location of his/her gaze corresponds to the symbol currently being processed in working memory [JUS76] and, moreover, that the eye naturally focuses on areas that are most likely to be informative [MAC70]. Eye-tracking will be employed in our work to help identify how gaze disparity in eye-location is affected at the network-level. By continuously monitoring user focus we aim to gain a better understanding why people do not notice obvious cues in the experimental video material. Eye-tracking will be measured at the network-level, however, due to the complexity of eye-tracking data, will be analysed separately to QoP data. Although the impact of delay and jitter on user perception of multimedia has been considered by other authors, these studies fail to considering both level of user understanding (information assimilation) and user satisfaction (both of the video QoS and concerning the user level of enjoyment).

- **Objective 2: Measurement of the perceptual impact of media-level parameter variation.** The *Human Visual System* (HVS) can only process detailed information within a small area at the centre of vision, with rapid acuity drop-off in peripheral areas [MAC70]. Attentive displays monitor and / or predict user gaze, to manipulate allocation of bandwidth, such that quality is improved around the viewer's point of gaze [BAR96]. Attentive displays offer considerable potential for the reduction of network resources and facilitate media-level quality variation with respect to both video content-based (technical-perspective) and user eye tracking-based (user-perspective) data. Accordingly, to consider media-level parameter variation, with consideration of both technical- and user-perspectives, we intend to measure the perceptual impact of using an attentive display system that manipulates frame-rate in Regions of Interest (RoI), where RoI areas are defined by analysing both video content- and user eye tracking-based data. Eye-tracking data will be used at the media-level to manipulate video content, yet no monitoring of user eye-gaze location will be made at the media-level.

- **Objective 3: Measurement of the perceptual impact of content-level parameter variation.** To consider user-perspective content-level parameter variation, we intend to measure the impact of display-type on user perception of multimedia quality. Devices used include a fixed head-position eye-tracker, a traditional desktop limited mobility monitor, a head-mounted display, and a personal digital assistant. These devices represent considerable variation in screen-size, level of immersion, as well as level of mobility, which are all of particular importance in the fields of virtual reality and mobile communications. Technical-perspective content-level parameter variation is achieved through use of diverse experimental video material. Although an eye-tracker will be used as a display device, no monitoring of user eye-gaze location data will be made at the content-level.

The structure of this document is as follows. In Chapter 2 we discuss issues relating to human perception of multimedia and how perception of multimedia quality has been defined across different studies, with the intent of defining research aims and objectives. In Chapter 3, we describe the research methodology that shall be used in experimental chapters 4, 5 and 6. In chapter 4, we simulate delay and jitter variation in order to measure the impact of network-level technical-perspective parameter variation on user perception of multimedia quality. Moreover, we independently monitor and analyse eye position, in order to provide a better understanding of the role that the human plays in the reception, analysis and synthesis of multimedia data at the network-level. In chapter 5 we manipulate media-level technical- and user-perspective parameters, in order to measure the subsequent change in user perception. We implement a novel frame rate based attentive display, which defines RoI areas from both video content-based (technical-perspective) and eye tracking data (user-perspective). Eye-tracking data will be used at the media-level to manipulate video content, yet no monitoring of user eye-gaze location will be made at the media-level. In chapter 6, we consider user-perspective content-level parameter variation, by measuring the impact of display-type on user perception of multimedia quality. Technical-perspective content-level parameter variation is achieved through use of diverse experimental video material. Although an eye-tracker will be used as a display device, no monitoring of user eye-gaze location data will be made at the content-level.Finally, in chapter 7, our research contributions are stated and experimental findings and considered.

CHAPTER 2

Defining the User-Perspective

Distributed multimedia quality has not traditionally been defined using a "single monotone dimension" [VIR95], but is a term that often means different things to different people [WAS97]. Indeed, a user's perception of distributed multimedia quality may be affected by numerous factors, e.g. delay or loss of frames, audio clarity, lip synchronisation during speech, video content, display size, resolution, brightness contrast, sharpness, colourfulness, as well as naturalness of video and audio content. [AHN93; APT95; KLE93; MAR96; ROU92].

In this chapter we introduce issues relating to human perception of multimedia and how perception of multimedia quality has been defined across different studies, with the intent of defining research aims and objectives. In section 2.1 we introduce four multimedia senses - Olfactory (smell), Tactile/Haptic (touch), Visual (sight) and Auditory (sound) – as they are closely linked to human perception. In section 2.2 we consider methods of quality definition, followed in section 2.3 by work relating to how each of the multimedia senses impacts user perception and ultimately user definition of multimedia quality. Special focus is given to video and audio information, as these are the senses on which our study focuses. Finally, in section 2.4, we compare a number of perceptual studies and accordingly state our research aims and objectives.

2.1. The Human Multimedia Senses

Our sensory systems acquire information about the current state of the world by gathering signals from receptors in the eyes, ears, and other sense organs. The signals from one side of the body are sent through nerve fibres to the cerebral cortex on the opposite side of the brain, where they are perceived and interpreted in terms of our previous experiences, knowledge, and expectations. In this section we aim to introduce the reader to the four physiological systems - *Olfactory* (smell), *tactile/haptic* (touch), *visual* (sight) and *auditory* (sound) – that lie at the core of the human perceptual / sensory experience.

2.1.1. The Human Olfactory System

The sense of smell or olfaction is used by humans as a means of identifying food, producing warnings of danger (e.g. rotten food, chemical dangers, fires), identify mates and predators, aiding navigation, as well as providing sensual pleasure [LEFF01]. Smell is important to the perception of an environment, and since olfactory neurons are directly connected to the brain, can directly influence our mood, can trigger discomfort, sympathy or even refusal [AXE95; MAM02].

Olfaction facilitates approximately 50 million primary sensory receptor cells in a small (2.5 cm^2) area of the nasal passage called the *olfactory region*. The olfactory region is formed of *cilia* projecting down out of the *olfactory epithelium* into a layer of mucous, which helps to transfer soluble odorant molecules to the receptor neurons. Above the olfactory epithelium, the neuronal cells form axons, which penetrate the cribriform plate of bone, thus reaching the olfactory bulb of the brain. Messages are sent directly to the higher levels of the central nervous system, via the olfactory tract, where olfactory information is decoded and a reaction is determined [LEFF01].

2.1.2. The Human Sense of Touch

Information from our skin allows us to identify several distinct types of sensations, as human skin containing a number of different sensory receptor cells that respond preferentially to various mechanical, thermal or chemical stimuli.

The majority of multimedia studies involve the tactile or touch sense, which detects pressure and touch (brushing, vibration, flutter and indentation), however, human skin is also sensitive to temperature and pain. There are five main types of sensory receptors, shown in Table 2.1 [MUR04], which react to different stimuli and with different adaptation speeds. Slowly adapting receptors send impulses to the brain when a constant stimulus is applied; in contrast, rapidly adapting receptors fire when stimuli is applied, yet do not send impulses as a result of constant stimulation.

Table 2.1: Characteristic of Sensory Receptors in the Skin [MUR04].

Receptor	**Stimuli**	**Sensation**	**Adaptation**
Merkel's disk	Steady indentation	Pressure	Slow
Meissner's corpuscle	Low frequency vibration	Gentle fluttering	Rapid
Ruffini's corpuscle	Rapid indentation	Stretch	Slow
Pacinian corpuscle	Vibration	Virbration	Rapid
Hair Receptor	Hair deflection	Brushing	Rapid or Slow

Information from the each skin receptor is carried along the "touch-neuron pathway" to the somatosensory cortex, which maps the senses in the body and transmits messages about sensory information to other parts of the brain, e.g. for use in performing actions, for making decisions, enjoying sensation or reflecting on them.

Sensory maps, in the cerebral cortex, are not uniformly distributed. Instead maps are defined by the density of sense receptors, which in turn reflects the importance of a particular body area at gathering tactile information. The homunculus (meaning 'little man'), shown in Figure 2.1, is a graphical representation of how the brain perceives the distribution of tactile senses in the human body.

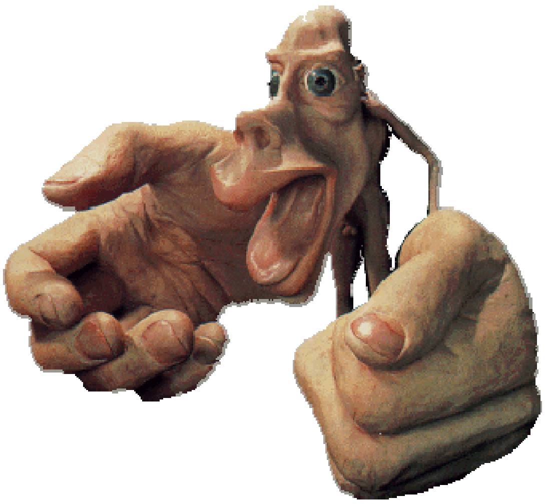


Figure 2.1: The Homunculus: Representing Sensory maps in the Cerebral Cortex [FAS04].

2.1.3. The Human Visual System

Light reflected from objects in the *visual field* enter the eye through the pupil and passes through the lens, which projects an inverted image of the object onto the *retina* at the back of the eye (see Figure 2.2 and Figure 2.3). The retina consists of approximately 127 million light-sensitive cells (120 million are called *rods*; 7 million are called *cones*, which can be subdivided into *L-cones, M-cones* and *S-cones*). Although cones are less light sensitive than rods, they are responsible for capturing colour within the human visual system.

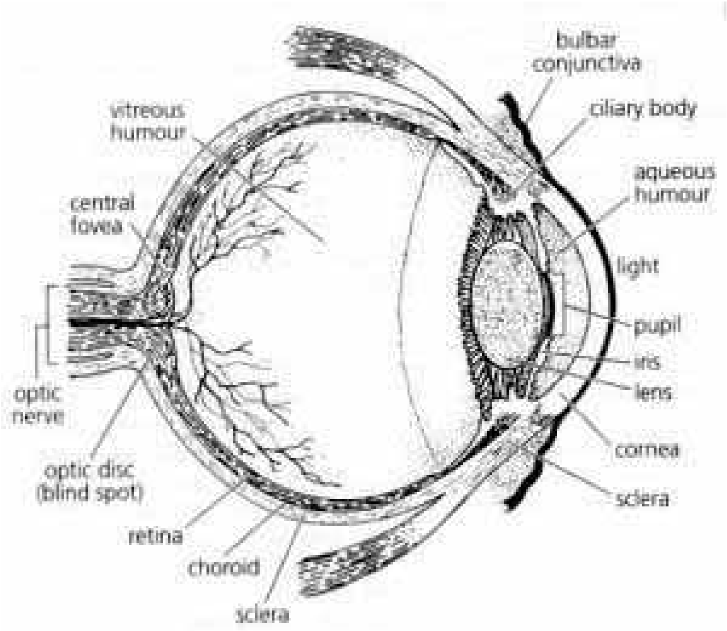


Figure 2.2: Diagram of the Human Eye [PRY04].

When light enters the eye, it must pass through seven cell-layers before reaching the rods and cones at the back of the eye (see Figure 2.3). The cells that process and transmit information to the brain are called the *bipolar, horizontal* and *ganglion* cells [MOL97].

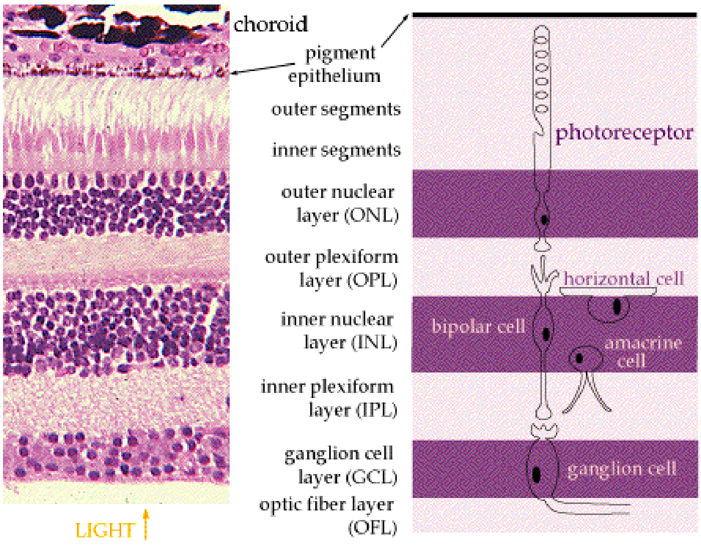


Figure 2.3: The Structure of the Human Retina [MOL97].

Photoreceptors at the back of the eye (cones and rods) are activated when light is shined at them, which consecutively activates bipolar cells. Importantly the output of a ganglion cell is only stimulated when certain bipolar activity occurs. Kuffler [KUF53] made electrophysiological measurements of the responses of retinal ganglion cells in cats and found that each ganglion cell took its input from a spatially localized region of the retina which he termed as its *receptive field* [KUF53]. Hartline, Wagner and Ratliff [HAT56] used small spots of light to stimulate specific parts of the receptive field and showed that, when a bipolar cell is activated, it strongly *inhibits* the output of neighbouring bipolar cells through a process call lateral inhibition. If a bipolar cell is activated then the output of the neighbouring bipolar cells is inhibited with the rate rising with the intensity of the light, and decreasing with the distance between neighbouring cells. This reaction is important as a ganglion cell is said to have a *centre-surround receptive field*, and can only be activated when the central area is light and the surrounding is mainly dark (see Figure 2.4).

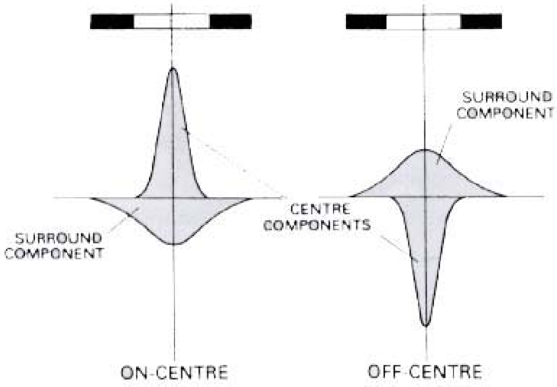


Figure 2.4: Enroth Cugel and Robon's (1966) model of the Centre-Surround Receptive Field [BRU96d].

Accordingly, instead of reacting to light level, ganglion cells react to either a small spot of light, a small ring of light, or a light-dark edge. If we introduce defused light across the whole receptive field then all bipolar cells are activated equally, causing all cells to be universally inhibited, which results in no change at the output of the ganglion cell. If a specific bipolar cell is activated, yet none of its neighbour cells are activated, then there will be no inhibition, which leaves the ganglion cell to fire at a faster than normal rate. When all neighbouring bipolar cells are activated, considerable inhibition exists. Consequently, the central ganglion cell shuts down and will not produce an output until light is turned off [BRU96a].

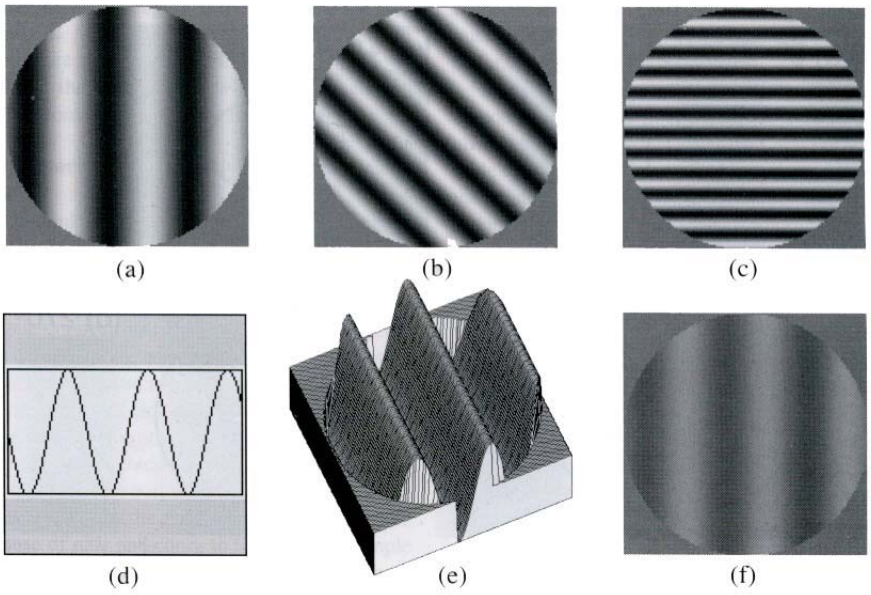


Figure 2.5: Sinusoidal grating vary in *orientation* and *spatial frequency* from a low (a) through medium (b) to high frequencies (c). The luminance profile (d) is a graph of the intensities against position, and in this example is a sine wave. In general, image intensity varies in two dimensions (x,y), and can be visualised as a surface (e). The grating shown in (f) has the same frequency as (a) but has a lower contrast. The luminance (or intensity) (L) of a sinusoidal grating as a function of position (x) is: L(x) = Lo + C + sin(2πfx - Φ). Lo is the mean luminance and f is the frequency of the grating. Contrast (C), ranging from –1 to +1, controls the amount of luminance variation around Lo. Changes in phase (Φ) shift the position of the grating along the x axis [BRU96b].

Enroth-Cugell and Robson [ENR96] used *sinusoidal gratings* (see Figure 2.5) as stimuli to identify the impact of waveform input on the retinal response of cats. Enroth-Cugell and Robson expected that:

i) when the peak of the grating, the part of maximum intensity, falls on the central bipolar cell, a increase in ganglion activity should occur as no neighbour inhibition exists (see Figure 2.6a).

ii) with a phase shift of 180°, there should be a decrease in ganglion response as the surrounding bipolar cells are active, thus inhibiting the ganglion activity (see Figure 2.6b).

iii) with a 90° phase shift, in either direction, no net difference exists between the centre of the central bipolar cell and its surrounding area, resulting is no change in ganglion activity (see Figure 2.6c and Figure 2.6d).

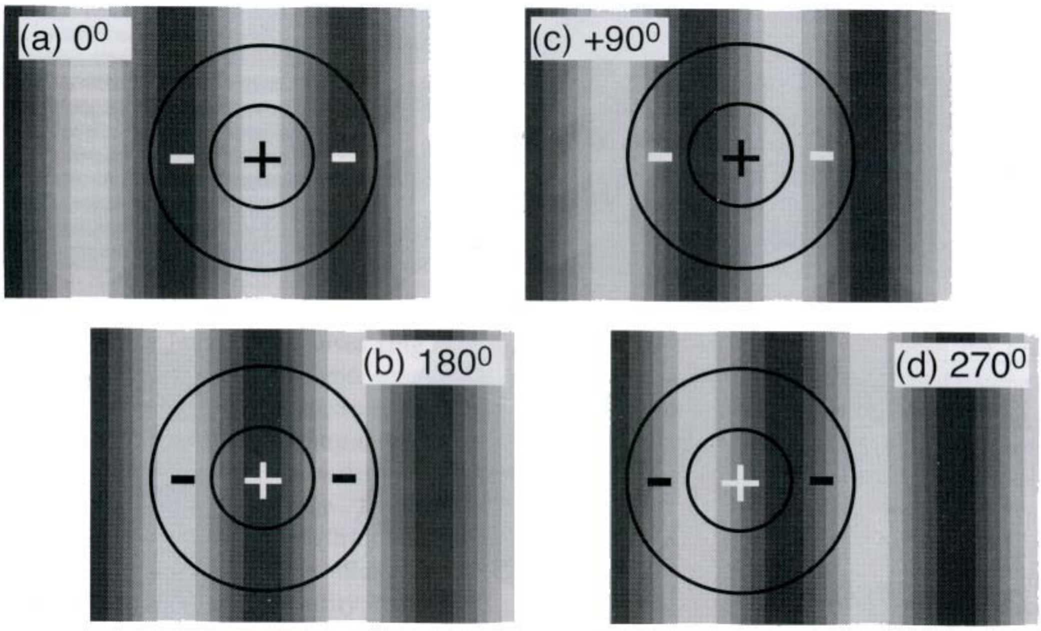


Figure 2.6: Phase shifts across Centre-Surround Receptive Field [BRU96c].

Interestingly, Enroth-Cugell and Robson showed that, for cats, some ganglion cells, which were termed *X cells*, did function as expected. However, in cats other ganglions cells, which are termed *Y cells*, were identified as having a non-linear response. de Monasterio and Gouras [DEM75] looked specifically at the retinas of primate monkeys, the closest genetic neighbour to humans, and showed that the majority of monkey ganglion cells were X type cells. X cell type posses a linear concentric ON / OFF receptive field, which consequently support the modelling of human vision. Interestingly, for primate vision X-type cells fall into two distinct categories: *P and M cells*. Both have centre and surrounding areas, however, unlike M cells, the two regions in a P cell are sensitive at different wavelengths supporting colour vision. Visual information from overlapping L-, M-, and S- cones are used by P ganglion cells to form new signals called *opponent colours*, identified by Hering [HER78], which facilitate human colour vision, define human colour acuity and support the use of opponent colours in optics.

If cones were distributed evenly across the retina, their average distance apart would be relatively large, and the ability to detect fine spatial patterns (*acuity*) would be relatively poor. Cones are therefore concentrated in the centre of the retina, in a circular area called *macula lutea*. Within this area, there is a depression called the *fovea*, which consists almost entirely of cones, and it is through this area of high acuity, extending over just 2° of the visual field, that humans make their detailed observations of the world. The fovea extends over a limited area of just 2x2 degrees of the visual field (180° horizontally and 140° vertically). The remaining part of the retina offers peripheral vision, which is characterized by being of only 15-50% of the acuity of the fovea [JAC95].

Linear mapping of visual receptivity at different areas of vision is facilitated by the fact that human ganglion cells are primarily X cells. Rodieck and Stone mapped the spatial organizations of ON / OFF receptors and showed that the distribution of receptivity on a primary visual cortex neuron could be modelled using Gaussian functions [ROD65]. Accordingly, the receptive field of neurons in the primary visual cortex can be modelled using a *Gabor function*, which describes a visual signal in both frequency and spatial domains (see Figure 2.7).

Visual contrast is a measure of the relative variation in luminance. It has been shown that different areas of the human eye have a band pass spatial frequency with a peak frequency ranging from 2-8 cycles per degree (cpd) [DEV82], however, due to the non-uniform organization of the receptive fields, contrast sensitivity also varies across the visual field. More precisely, the eye detects a signal if, and only if, its contrast is greater than the detection threshold.

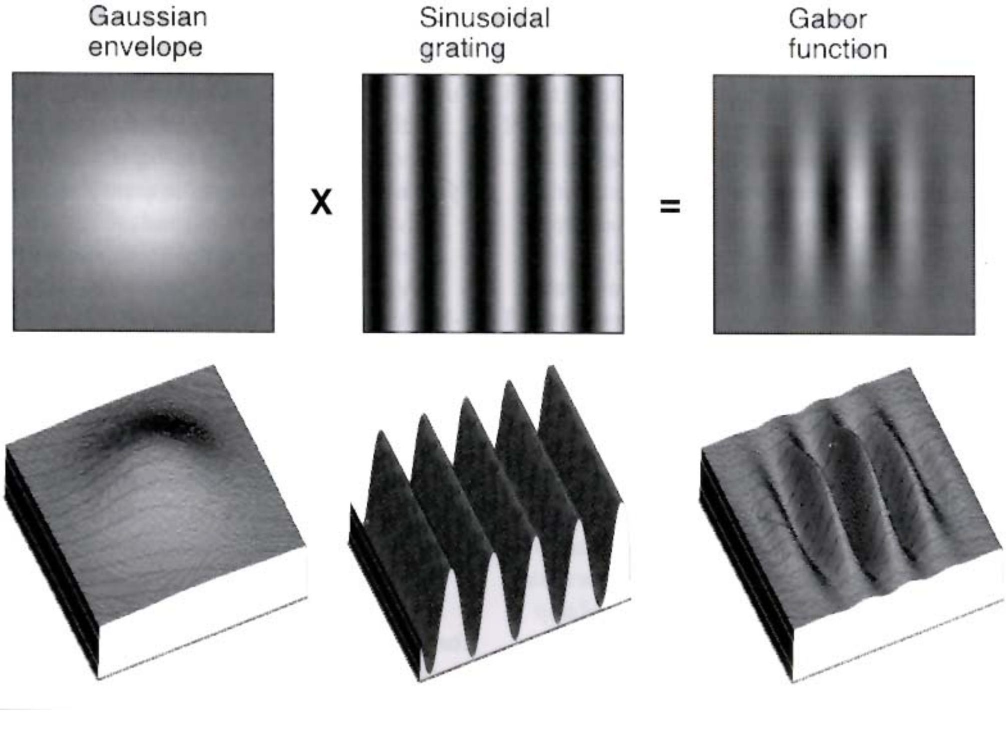


Figure 2.7: Receptive field of Neurons in the Primary Visual Cortex [BRU96e].

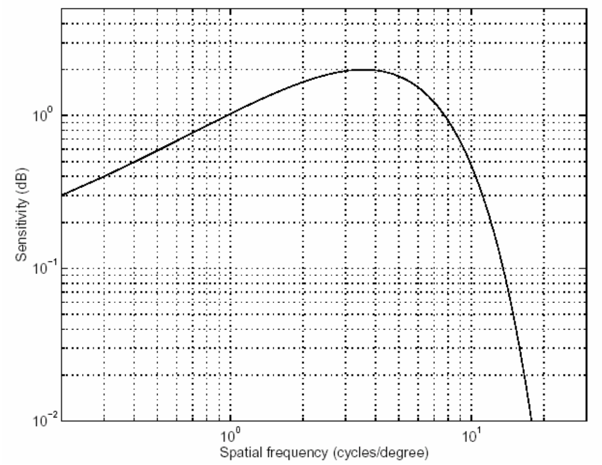


Figure 2.8: Illustration of the Sensitivity of the Human Eye as a Function of Spatial Frequency [DEV82].

The inverse of this detection threshold is called the *contrast sensitivity function* (CSF) and defines contrast sensitivity in terms of frequency. A typical CSF is illustrated in Figure 2.8. More generally, the sensitivity of the eye varies as a function of *spatial frequency*, *orientation* and *temporal frequency*. Whilst contrast sensitivity deals with the visibility of a single stimulus, *contrast masking* accounts for interferences between stimuli. Masking occurs when a stimulus that is visible by itself cannot be detected due to the presence of another. Facilitation, the opposite effect, occurs when a stimulus that is not visible by itself can be detected due to the presence of another. Masking explains why similar coding artefacts are obvious in certain regions of an image, yet are hardly noticeable elsewhere.

Masking has been studied and it has been shown that its effect is maximal when the stimulus and the masker have similar orientation, spatial and temporal frequencies, and decrease as the distance between signals increases. A common model of masking is a non-linear transducer as shown in Figure 2.9b. Consider two stimuli, the "signal" and the "masker". CTO defines the detection threshold of a signal in the absence of the masker. CT defines the detection threshold of the signal measured in the presence of a masker, whilst CM defines the contrast of the masker. Three regions are identified:

- At low values of CM, the detection threshold remains constant (CT = CTO).
- As CM gets closer to CTO, the detection threshold slightly decreases, as show in Figure 2.9a.

- Finally, as CM increases, CT increases as power of the contrast masker and is therefore linear in a log-log graph. The angle of the slope if denoted by ε.

Interestingly, as masking thresholds and Gabor functions can be modelled, they are fundamental to computational models of human vision.

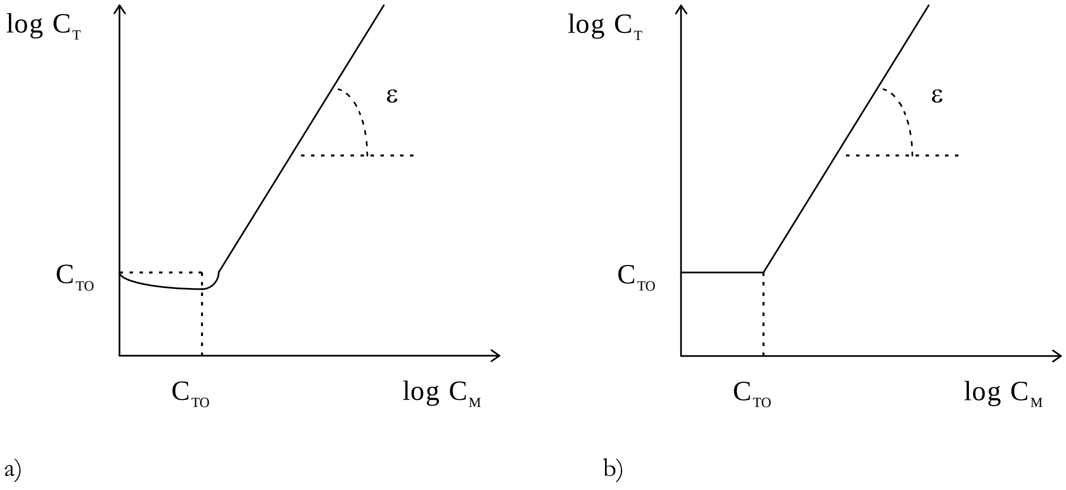
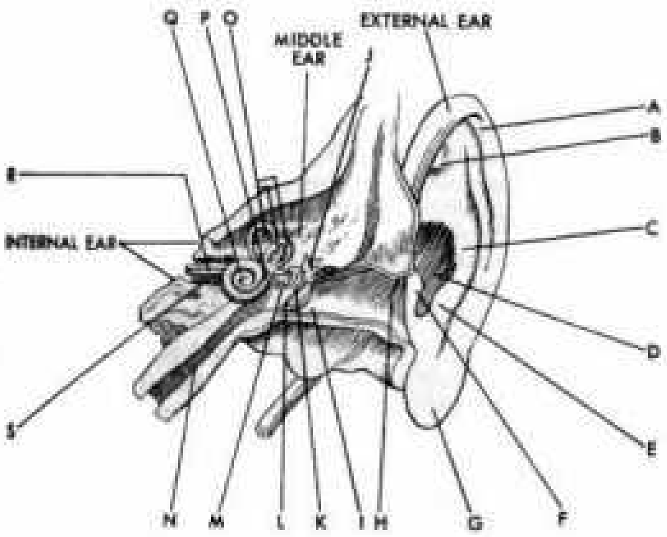


Figure 2.9: Model of the masking phenomenon [GHP00].

2.1.4. The Human Auditory System

When an object vibrates, it produces a sequence of wave compressions in the air surrounding it. These fluctuations in air pressure spread away from the source of vibration, reducing in magnitude as the energy is dispersed. When any two or more waveforms occur at the same time, they interact creating a new waveform that is the sum of its component parts.

Sound is simply the sensation produced by the ear when a vibration of an object occurs within the frequency range (20 Hz to 20 KHz) audible to humans. The volume of sound, at the source of vibration, is dependant upon the magnitude of sound energy being produced. The frequency is dependent upon the frequency of compressions being produced by the source of vibration.



human ear (transverse section)
External ear: A, helix; B, fossa of antihelix; C, antihelix; D, concha; E, antitragus; F, tragus; G, lobe; H, ear canal. Middle ear: I, eardrum; J, malleus; K, incus; L, tympanic cavity; M, stapes; N, Eustachian tube. Internal ear: O, semicircular canals; P, vestibule; Q, cochlea; R, auditory nerve, S, internal auditory meatus

Figure 2.10: Diagram of Human Ear [PRA04].

The ear is divided into three parts - *the outer (external), the middle* and the *inner (internal) ear* (see Figure 2.10). The outer ear collects sound waves and focuses them along the ear canal to the eardrum. The eardrum vibrates, causing bones (*Malleus and Incus*) to rock back and forth, which passes movement to the cochlea where fluid in the inner ear is disturbed (see Figure 2.10). The disturbance of fluid causes thousands of small hair cells to vibrate. The cochlea converts sound waves into electrical impulses, which are passed on to the brain via the auditory nerve [HEA04].

Hearing is mainly processed in the *temporal lobe* (at the side of head above the ears), which is responsible for primary organization of sensory inputs [REA81]. The temporal lobe is also responsible for memory acquisition, auditory sensation and perception, organization and categorization of verbal material, long-term memory, affective behaviour and some sexual behaviour [KOL90].

In multimedia, an accurate understanding of the workings of the human auditory system is indispensable due to the heavy use of sound, speech, music and special effects. Further more, redundancies and limitation of both the human audio and visual systems are exploited in digital compression techniques, which is closely linked to perceptive quality assessment issues, an issue that we will now address.

2.2. Quality Assessment

There are two main approaches to quality assessment: *subjective* (which uses human viewers) and *objective* (via appropriately defined quality metrics). Any information that derives from a person can be considered as being 'subjective' in nature. Any information that is recorded without chance of bias is considered as being 'objective' [MULL01]. Subjective data can be either structured (i.e. questionnaires and checklists) or unstructured in nature (i.e. open interviewing or participant observation) and can be designed to produce either rich or concise data, depending on desired interpretation. Objective data ensures dependability, however interpretation is limited by experimental design. In this section we consider methods of assessing what is perceived to be "quality".

2.2.1. Subjective Testing

Multimedia applications are produced for the enjoyment / and or education of human viewers, so their opinion of the presentation quality is important to any quality definition. This is why subjective quality ratings form the benchmark of quality definition.

2.2.1.1. Single Stimulus Methods

Single Stimulus (SS) methods are used when multiple separate scenes are shown. There are two approaches: SS with no repetition of test scenes and SSMR (Single Stimulus with Multiple Repetition) where the test scenes are repeated multiple times. Different measurement scales include:

1. *Quality Scale* – this is a single stimulus quality scale commonly using a five point category where subjects assess the quality of the material giving scores from 5 to 1, corresponding to *excellent, good, fair, poor,* and *bad*, respectively. It is recommended by the International Telecommunications Union for appraisal of both audio (the ITU-T recommendations [ITU80]) and video quality (the ITU-R recommendations [ITU50]). For example: listeners rating the quality of specific test sentences, where each listener gives every sentence a rating as

follows: (1) bad; (2) poor; (3) fair; (4) good; (5) excellent. The *Mean Opinion Score* (MOS) is then the arithmetic mean of all the individual scores, and can range from 1 (worst) to 5 (best).

2. *Adjectival* (or *impairment*) *Scale* – this is a single stimulus scale that measures feedback on an overall impression scale of impairment: *imperceptible, perceptible but not annoying, slightly annoying, annoying,* and *very annoying*; however, half-grades may be allowed. Where tests involve only audio, rating can also be done on a listening effort scale where the grade might range from '*complete understanding: no effort required*' to '*misunderstood without reasonable effort*'.

3. *Binary Scale* - A single stimulus binary scale might be used with the subject answering 'yes' or 'no'. For instance, after grading an application using the quality scale, users might be asked whether or not they had any difficulty using it.

4. *Numerical Scale* - This is commonly a 7- or 11- grade numerical scale, useful if a reference is not available.

5. *Non-categorical scale* - a continuous scale with no numbers and / or a large range, e.g. 0 - 100.

6. *Single Stimulus Continuous Quality* (SSCQE) - Instead of viewing stimuli of limited duration, participants watch a program that has been manipulated as a result of test conditions. Using a slider, the subjects continuously rate the instantaneous perceived quality on a scale from bad to excellent [BOU98].

2.2.1.2. Double Stimulus Methods

Here viewers are shown both test material as well as material depicting reference conditions. The main types of scales used are:

1. *Double Stimulus Impairment Scale* (DSIS) - Observers are shown multiple reference-scene / degraded scene pairs. The reference scene is always first. Scoring is made using an impairment scale (see above).

2. *Double Stimulus Continuous Quality Scale* (DSCQS) - Here observers are shown multiple scene pairs with the reference and degraded scenes randomly ordered,

i.e.: the user is not told which of the two is the original. The observer is then asked to rate the quality of the two by drawing a mark on a continuous quality scale that ranges from excellent to bad. Each scene in the pair is separately rated, however in reference to the other scene in the pair. Analysis is based on the difference in rating, rather than the absolute values.

3. *Stimulus Comparison Method* - is usually accomplished with two well-matched monitors but may be done with one. The differences between scene pairs are scored in one of two ways:

    - Adjectival - a 7-grade, +3 to -3 scale labelled: *much better*, *better, slightly better, the same, slightly worse, worse*, and *much worse.*

    - Non-categorical - a continuous scale with no numbers or a relation-number either in absolute terms or related to a standard pair.

Subjective testing methods have different applications. Single stimulus subjective tests are used when a single scene is shown, though because of limitations of human working memory [ALD95], SSCQE is used if clip duration is substantially longer than 20-30 seconds. DSIS is used when identifying clear visible errors between two stimuli, whilst DSCQ is used when test and reference stimuli are similar, as DSCQ is sensitive to small differences.

2.2.2. Objective Testing

Any information that is recorded without chance of bias is considered as being ‘objective’ in nature [MULL01]. Objective testing can therefore be used independent of the underlying transmission system, experimental conditions or the encoding process. Accordingly objective tests provide an excellent means of quality standardisation and comparison.

2.2.2.1. Signal to Noise Ratio

Noise can originate from many different sources and is considered to be any part of the signal that does not represent the parameter being measured. In analogue and digital communications, the signal-to-noise ratio, often written S/N or SNR, is the measure of signal strength relative to background noise.

$$\text{SNR} = 20 \log_{10}(\text{Signal/Noise}) \ \textit{(dB)}$$

A low noise signal has a high SNR, while a high noise signal has a low SNR. The metric is usually expressed in decibels (dB) and in terms of peak values for impulse noise and root-mean-square (RMS) values for random noise. If peak values are used for the magnitude of the amplitude of the signal then the metric becomes the *Peak Signal-to-Noise-Ratio* (*PSNR*).

2.2.2.2. Video Quality Metrics

Traditional analogue techniques for measuring video quality, e.g. RMS and PSNR, have shown considerable limitations when used in the digital paradigm. Although useful for comparison, analogue techniques consistently fail to simulate user perceived quality, due to digital artefacts. Digital artefacts includes blockiness, blurring, ringing, colour bleeding and motion mismatches [YEU98], with the perceptual impact of artefacts being dependent on the image content. Content-dependent error limits the effectiveness of traditional objective quality measures, which operated solely on a pixel-to-pixel basis. Pixel-to-pixel analysis provides a measurable comparison of error, yet neglects to show the important influence of image content and viewing condition on a user's overall perception of error.

Quality is ultimately defined from the user-perspective, yet subjective measurement of user perception is both time consuming and financially costly. To avoid subjective evaluation - thus limiting the inefficient use of time, as well as a considerable expense - methods of objective video quality assessment have been developed for use with digital video. Whilst systems exist for assessing the quality of still images (*see* [AHA93] *for a review*), we are more interested with their extension to moving pictures. Accordingly, this section aims to show the reader how spatial- temporal- resolution, spatial- temporal- masking and cognitive processes have been used to produce objective assessment of video quality.

Video-based quality metrics were first used by Lukas and Budrikis [LUK82], who proposed a metric using a *spatio-temporal model* of the human visual system. Although other metrics were later developed [MAT91; WAT90], it is only in recent years that there has been an increased interest in perceptual video quality assessment [CAR98; HOR96; LOD96; TAN98; VAN96; VAV96; WAT98; WES97; WIN99].

Perceptual video quality assessment methods can be classified along many axes, including: resolution, masking and cognitive processes.

***Resolution*:** Early resolution models measured a single frequency, a single orientation and did not include consideration of temporal mechanisms [LUK82; MAN74; SCH56].

- Spatial considerations for both chromatic and achromatic data were incorporated by Teo and Heeger, and Winkler [TEO94; WIN98].

- Temporal mechanisms are included in [FRE97; FRE98].

- Front end filters, or *filter banks* can be used over a range of frequencies and orientations. Pyramid structures are used in a number of studies including Watson's cortex transform [WAT87], which was later modified by Daly [DAL93]. Simoncelli et al. proposed the *steerable pyramid* [SIM92; SIM95] and the *Quadrature Mirror Filter* (QMF) transform [SIM90], which was used for the perceptual distortion measurement by Teo and Heeger [TEO94].

***Masking*:** Masking is important as it describes the interaction between different stimuli. Masking has been shown to occur between different orientations [FOL94], between different spatial frequencies, and luminance channels [COL90; LOS94; SWI88]. Accordingly, spatial masking has been used to model the inhibitory effect of the eye [WAT97], as well as a direct model of neural cell responses [TEO94]. Temporal masking is a sudden rise in visibility, as a result of a sudden scene change. Temporal masking is less developed than spatial masking, yet has been shown to be useful by Girod [GIR89] and Watson [WAT98].

***Cognitive Processes*:** Cognitive behaviour varies greatly between individuals. However, two important components have been used to produce models of human vision i) focus of attention and ii) object tracking.

- Focus of Attention: When viewing a video, observers focus gaze on particular areas [END94; STT91; STT94], showing that viewing is not participant dependent. Yarbus showed that during visual perception and recognition, human eyes move and successively fixate the most informative parts of the image [YAR67]. Accordingly, focus of interest is highly dependent on both scene content and current task [KAH73]. Accordingly, Maeder et al. [MAE96] and subsequently Osberger et al. [OSB97; OSB98] proposed the prediction of user attention by accounting for perceptual factors, such as edge strength, texture energy, contrast, colour variation and homogeneity, which have been identified as essential for the definition of objects in human visual [COH90; FIN80; HUB70; MOR93; NIE97; OSB97; YAR67].

Object Tracking: Viewers naturally track movement in the visual field [HIL94]. Tracking movement is essential because of the spatial acuity characteristics of the human visual system, which is dependent on reducing the velocity of the object image on the retina. Smooth pursuit is used to track a moving object, which works well even at high velocities, but is impeded by large acceleration or unpredictable motion [ECK93]. Interestingly, although tracking objects increases object acuity, tracking a particular object using smooth pursuit reduces the spatial acuity of objects in the background area.

2.3. Multimedia Senses – Perceptual Implications.

In this section we consider work relating to each of the multimedia senses and how they impact user perception and ultimately user quality definition.

2.3.1. Olfactory

Research in the field of olfaction is limited, as there is no consistent method of testing user capability of smell. The first smell based multimedia environment (sensorama) was developed by Heilig [HEI62; HEI92], which simulated a motorcycle ride through New York and included colour 3D visual stimuli, stereo sound, aroma, and tactile impacts (wind from fans, and a seat that vibrated).

A major area of olfactory research has been to explore whether scent can be recorded and therefore replayed to aid olfactory perceptual displays [DAV01; RYA01; NA02]. As a result, Cater [CAT92; CAT94] successfully developed a wearable olfactory display system for a fire fighters training simulation with a Virtual Reality (VR) orientated olfactory interface controlled according to the users location and posture. In addition, researchers have used olfaction to investigate the effects of smell on a participant’s sense of presence in a virtual environment and on their memory of landmarks. Dinh et al. [DIN99] showed that addition of tactile, olfactory and /or auditory cues within a virtual environment increased the user's sense of presence and memory of the environment.

Unlike all other senses, results from olfaction studies suggest that there are cultural differences in odour perception. Ayabe-kanamura et al. [AYA98] tested groups of Japanese and German subjects for their odour perceptions of typical Japanese and German dishes (e.g. sushi and beer). Their results indicated that cultural backgrounds lead to differences in odour quality perception, however, no firm conclusions were made.

2.3.2. Tactile /Haptic

Current research in the field of haptics focuses mainly on either sensory substitution for the disabled (tactile pin arrays to convey visual information, vibrotactile displays for auditory information) or use of tactile displays for teleoperation (the remote control of robot manipulators) and virtual environments. Skin sensation is essential, especially when participating in any spatial manipulation and exploration tasks [HOW02]. Accordingly, a number of *tactile display* devices have been developed that simulate sensations of contact. Whilst "Tactile display" describes any apparatus that provides haptic feedback, tactile displays can be subdivided into the follow groups:

- ***Vibration*** sensations can be used to relay information about phenomena, such as surface texture, slip, impact, and puncture [HOW02]. Vibration is experienced as a general, non-localised experience, and can therefore be simulated by a single vibration point for each finger or region of skin, with an oscillating frequency range between 3 and 300 Hz [KON95; MIN96].

- ***Small-scale shape or pressure*** distribution information is more difficult to convey than that of vibration. The most commonly used approach is to implement an array of closely aligned pins that can be individually raised and lowered against the finger tip to approximate the desired shape. To match human finger movement, an adjustment frequency of 0 to 36 Hz is required, and to match human perceptual resolution, pin spacing should be less than a few millimetres [COH92; HAS93; HOW95].

- ***Thermal displays*** are a relatively new addition to the field of haptic research. Human fingertips are commonly warmer than the "room temperature".

Therefore, thermal perception of objects in the environment is based on a combination of thermal conductivity, thermal capacity, and temperature. Using this information allows humans to infer the material composition of surfaces as well as temperature difference. A few thermal display devices have been developed in recent years that are based on Peltier thermoelectric coolers - solid-state devices that act as a heat pump, depending on direction of current. [CAL93; INO93].

Many other tactile display modalities have been demonstrated, including *electrorheological devices* (a liquid that changes viscosity electroactively) [MON92], *electrocutaneous stimulators* (that covert visual information into a pattern of vibrations or electrical charges on the skin), *ultrasonic friction displays*, and *rotating disks* for creating slip sensations [MUR04].

2.3.3. Sight and Sound

Although the quality of video and audio is commonly measured separately, considerable work shows that audio and video information is symbiotic in nature, that is one medium can have a impact on the user's perception of the other [WAS96; RIM98]. Moreover, the majority of user multimedia experience is based on both visual and auditory information. As a symbiotic relationship has been shown between the perception of audio and video media, we consider multimedia studies concerning the variation and perception of sight and sound together.

2.3.3.1. Modelling Distributed Multimedia

There are numerous factors that have been shown to influence distributed multimedia video quality, e.g. delay or loss of frames, audio clarity, lip synchronisation during speech, as well as the general relationship between visual auditory components [APT95]. As a result, considerable work has been done looking at different aspects of distributed multimedia video quality at many different levels. Unfortunately, as a result of multiple influences on user perception of multimedia quality, providing a succinct, yet extensive review of such work is complex. To aid this review, and ultimately the definition of research aims and objectives, we extend a quality model, which was first used by Wikstrand [WIK03].

Wikstrand segregates quality into three discrete levels: the *network-level*, the *media-level* and *content-level*.

- The network-level is concerned with how data is communicated over the network and includes variation and measurement of parameters including: bandwidth, delay, jitter and loss.
- The media-level is concerned with how the media is coded for the transport of information over the network and / or whether the user perceives the video as being of good or bad quality. Media-level parameters include: frame rate, bit rate, screen resolution, colour depth and compression techniques.
- The content-level is concerned with the transfer of information and level of satisfaction between the video media and the user, i.e. level of enjoyment, ability to perform a defined task, or the user's assimilate critical information from a multimedia presentation.

Wikstrand showed that all factors that influence distributed multimedia quality can be categorised by assessing the information abstraction. The network-level concerns the transfer of data and all quality issues related to the flow of data around the network. The media-level concerns quality issues relating to the transference methods used to convert network data to perceptible media information, i.e. the video and audio media. The content-level concerns quality factors that influence how media information is perceived and understood by the end user.

At each quality abstraction defined in Wikstrand's model, quality parameters can be adapted, e.g. jitter at the network-level, frame rate at the media-level and finally display-type at the content-level. Similarly, at each level of the model, quality can be measured, e.g. percentage of loss at the network-level, user mean opinion score (MOS) at the media-level, and task performance at the content-level. In our work, and in addition to the model proposed by Wikstrand, we incorporated two distinct quality perspectives, which reflect the infotainment duality of multimedia: the user-perspective and the technical-perspective.

- ***User-Perspective***: The user-perspective concerns quality issues that rely on user feedback or interaction. The user-perspective can be adapted and measured at the media- and content-levels. The network-level does not facilitate the user-perspective since user perception can not be measured at this low level abstraction.

- ***Technical-Perspective*:** The technical-perspective concerns quality issues that relate to the technological factors involved in distributed multimedia. Technical parameters can be adapted and measured at all quality abstractions.

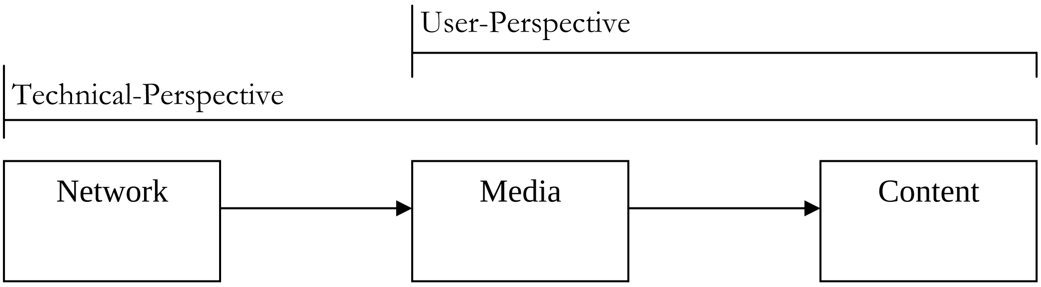


Figure 2.11: Quality Model, demonstrating Network-, Media- and Content-Level Abstractions and Technical- and User-Perspectives.

Figure 2.11 presents the model of distributed multimedia quality that will be used in our work to aid both a succinct yet extensive review of related research, yet also the definition of research aims and objectives.

The following sections describe perceptual studies, in context of the identified quality model (see Figure 2.11). Each of the following sections concerns a quality abstraction level (network-, media- or content-level) and includes work relating to studies that adapt and measure quality factors at the defined perspectives (technical-perspective or media-perspective). Each subsection, concerning quality variation and measurement, initially introduces the findings of all relevant studies, yet a detailed description of each new study is given at the end of the relevant subsection.

2.3.3.2. The Network-Level

Network-Level Quality Variation (Technical-perspective)

- Ghinea and Thomas [GHI00] manipulated bit error, segment loss, segment order, delay and jitter in order to test the impact of different media transport protocols on user perception understanding and satisfaction of a multimedia presentation.

- Claypool and Tanner [CLA99] manipulated jitter and packet loss to test the impact on user quality opinion scores.

- Procter et al. [PRO99] manipulated the network load to provoke degradations in media quality.

Ghinea and Thomas (2000): Ghinea and Thomas [GHI00] tested the impact on user Quality of Perception (QoP) of an adaptable protocol stacks geared towards human requirements (RAP - Reading Adaptable Protocol), in comparison to different media transport protocols (TCP/IP, UDP/IP). RAP incorporates a mapping between QoS parameters (bit error rate, segment loss, segment order, delay and jitter) and Quality of Perception (QoP) – the user understanding and satisfaction of a video presentation. Video material used included 12 windowed (352*288 pixel) MPEG-1 video clips, each between 26 and 45 seconds long, with a consistent objective sound quality, colour depth (8-bit) and frame rate (15 frames per second). The clips were chosen to cover a broad spectrum of subject matters, whilst also considering the dynamic, audio, video and textual content of the video clip.

Results showed that:

- RAP enhanced user understanding, especially if video clips are highly dynamic. TCP/IP can be used for relatively static clips. UDP/IP performs badly, in context of information assimilation.

- Use of RAP successfully improves user satisfaction (10/12 videos). TCP/IP received the lowest associated satisfaction ratings.

- RAP, which incorporates the QoS to QoP mapping [GHI00], is the only protocol stack used, which was not significantly different to those identified when video were shown on a standalone system. Accordingly, RAP effectively facilitates the provision of user QoP.

Claypool and Tanner (1999): Claypool and Tanner [CLA99] measured and compare the impact of jitter and packet loss on perceptual quality of distributed multimedia video.

Results showed that:

- Jitter can degrade video quality nearly as much as packet loss. Moreover, the presence of even low amounts of jitter or packet loss results in a severe degradation in perceptual quality. Interestingly, higher amounts of jitter and packet loss do not degrade perceptual quality proportionally.

- The perceived quality of low temporal aspect video, that is video with a small difference between frames, is not impacted as much in the presence of jitter as video with a high temporal aspect.

- There is a strong correlation between the average number of quality degradation events (points on the screen where quality is affected) and the average user quality rating recorded. This suggests that the number of degradation events is a good indicator of whether a user will like a video presentation effected by jitter and packet loss.

Procter et al. (1999): Procter et al. [PRO99] focuses on the influence of different media content and network Quality of Service (QoS) variation on a subject's memory of, and comprehension of, the video material. In addition, Procter et al. focus on the impact of degraded visual information on assimilation of non-verbal information.

A simulation network was used to facilitate QoS variation. Two 4 Mbs token rings were connected by a router, so that packets had to pass through the router before being received by the client. Two background traffic generators were used to generate two network conditions i) no load (with no traffic) and ii) load (with simulated traffic). During the load condition, packet loss was 'bursty' in character, varying between zero and one hundred percent, yet had an overall average between thirty to forty percent. Audio quality was subsequently dependent on network conditions. Video material used consisted of two diametrically opposed presentations: a bank's annual report and a dramatized scene of sexual harassment. Two experiments were used:

- The first experiment was designed to investigate the effects of network QoS on a subject's assessment of quality. Two questionnaires were used: the first concerned the subjective evaluation of quality (with scores 1-5 representing very poor and very good respectively) as well as factors that had impaired quality (caption quality, audio quality, video quality, audio/video synchronisation and gap in transmission); the second questionnaire measured a participants comprehension of the material based upon their recall of its factual content. Results showed that subjects rated the quality higher in the non-load than in the load condition, with overall impression, ease of understanding and technical quality being the significantly better quality with no network load. Two factors were found to significantly impair quality in the no-load network condition: video quality and audio/video synchronisation. When a network load was added, audio quality, video quality, audio/video synchronisation as well as transmission gaps were found to significantly impair user perception of multimedia quality. No difference was measured in the level of factual information assimilated by users.

- The second experiment investigated the impact of visual degradation of the visual channel on the uptake of non-verbal signals. Again, two network load conditions were used i) no load and ii) load, to simulate network traffic. The same test approach was used, however the second questionnaire considered a) factual questions, b) content questions relating to what they thought was happening in a dramatised section of the video, i.e. the participants ability to judge the emotional state of people, and c) questions asking the user to specify his/her confidence with his/her answers. Results showed that subjects rated quality higher in the no-load condition, with overall impression, content, ease of understanding and technical quality rating being significantly higher under no-load conditions. In the no-load condition audio, video quality and audio-video synchronisation were considered to have an effect on user perception of multimedia quality. In the load network condition, caption quality, audio quality, video quality, audio/video synchronisation and gap in transmission all were shown to have an impairing impact on user perception of multimedia quality. No significant difference was measured between the level of factual information assimilated by users when using load and non-load conditions. In conclusion, Procter et al. [PRO99] observed that degradation of QoS has a greater influence on a subjects' uptake of emotive / affective content than on their uptake of factual content.

Network-Level Quality Measurement (Technical-Perspective)

- Loss - Loss is a percentage measure of data packets (audio, video), which are dropped or 'lost' during a distributed multimedia presentation. Loss occurs due to network congestion and can therefore be used as an indication of end-to-end network performance or "quality" (Ghinea and Thomas [GHI00]; Koodli and Krishna, [KOO98]).

- Delay and Jitter - Delay is the time taken by a packet to travel from the sender to the recipient. A delay is always incurred when sending distributed video packets, however the delay of consecutive packets is rarely constant. The variation in the period of delay is called the jitter. Wang et al. [WAN01] used jitter as an objective measure of video quality.

- Bandwidth - The granularity of video information is defined by media-level technical factors including frame rate, bit rate as well as spatial and chromatic resolution. The end-to-end bandwidth is defined as the network resource that facilitates the provision of these media-level technical parameters. Accordingly, the available end-to-end bandwidth is important since it determines the network resource available to applications at the media-level. Wang et al. [WAN01] measured bandwidth impact for their study of real world media performance.

Koodli and Krishna (1998): Koodli and Krishna [KOO98] describe a metric called noticeable loss. Noticeable loss captures inter-packet loss patterns and can be used in source server and network buffers to pre-emptively discard packets based on the "distance" to the previous lost packet of the same media stream. Koodi and Krishna found that incorporation of noticeable loss greatly improves the overall QoS, especially in the case of variable bit rate video streams.

Wang, Claypool and Zuo (2001): Wang et al. [WAN01] presented a wide-scale empirical study of RealVideo traffic from several internet servers to many geographically diverse users. They found that when played over a best effort network, RealVideo has a relatively reasonable level of quality, achieving an average of 10 video frames per second and very smooth playback. Interestingly, very few videos achieve full-motion frame-rates. Wang et al. showed that: with low-bandwidth internet connection video performance is most influenced by the user's limited bandwidth connection speed; with high-bandwidth internet connection the performance reduction is due to server bottlenecks. Wang et al. [WAN01] used level of jitter and video frame rates as objective measures of video quality.

2.3.3.3. The Media-Level

Media-Level Quality Variation (Technical-Perspective)

- Apteker et al. [APT95], Ghinea and Thomas [GHI98], Kawalek [KAW95], Kies et al. [KIE97], Masry et al. [MAS01], Wilson and Sasse [WIL00a; WIL00b], and Wijesekera [WIJ99] manipulated video or audio frame rate.

- Gulliver and Ghinea [GULL03] manipulated use of captions.

- Wikstrand and Eriksson [WIK02] used various animation techniques to model football matches for use on mobile devices.

- Hollier and Voelcker [HOL97] varied audio-video quality to examine inter-stream reliance.

- Kies et al. [KIE97] manipulated image resolution.

- Steinmetz [STE96] and Wijesekera et al. [WIJ99] manipulated video skew between audio and video, i.e. the synchronisation between two media.

- Steinmetz [STE96] manipulated the synchronisation of video and pointer skews.

- Masry et al. [MAS01] and Winkler [WIN01] tested different video compression codecs.

Apteker et al. (1995): defined a *Video Classification Scheme* (VCS) [APT95] to classify video clips (see Table 2.2), based on three dimensions, considered inherent in video messages: the temporal (T) nature of the data, the importance of the auditory (A) components and the importance of visual (V) components.

Table 2.2: Video Classification Examples [APT95].

Category Number	**Video Information**	**Definition**
1	Logo/ Test Pattern	Tlo Alo Vlo
2	Snooker	Tlo Alo Vhi
3	Talk Show	Tlo Ahi Vlo
4	Stand-up Comedy	Tlo Ahi Vhi
5	Station Break	Thi Alo Vlo
6	Sporting Highlights	Thi Alo Vhi
7	Advertisements	Thi Ahi Vlo
8	Music Clip	Thi Ahi Vhi

"High temporal data" concerns video with rapid scene changes, such as general sport highlights, "Low temporal data" concerns video that is largely static in nature, such as a talk show. A video from each of the eight categories was shown to users in a windowed multitasking environment. Each multimedia video clip was presented in a randomised order at three different frame rates (15, 10, and 5 frames per second). The users then rated the quality of the multimedia videos on a 7 point graded scale. Apteker et al. showed that:

- Video clips with a lower video dependence (Vlo) were considered as more watchable than those with a high video dependence (Vhi).

- Video clips with a high level of temporal data (Thi) were rated as being more watchable than those with a low level of temporal data (Tlo).

- Frame-rate reduction itself leads to progressively lower ratings in terms of watchability.

- There exists a threshold, beyond which no improvement to multimedia quality can be perceived, despite an increase in available bandwidth, which is supported by [FUK97; GHP00; STE96; VAN96].

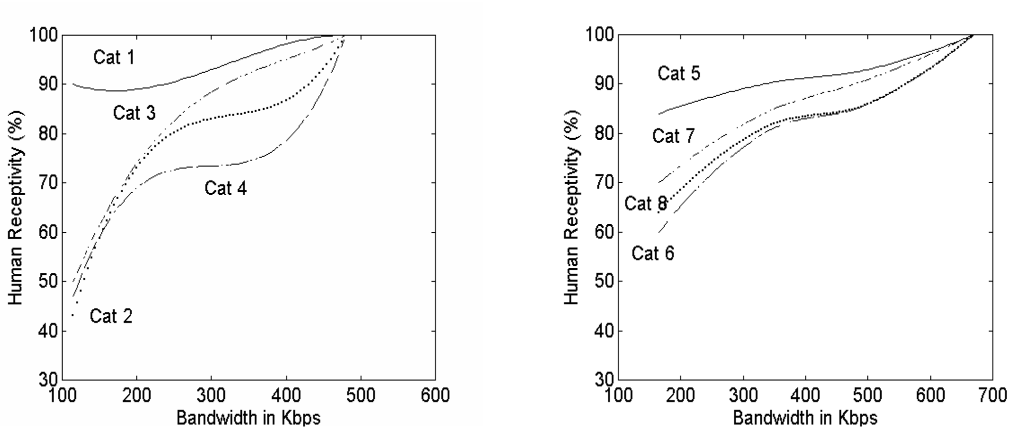


Figure 2.12: Human Receptivity vs. Bandwidth curves from [APT95].

Apteker et al. expressed human receptivity as a percentage measure, with 100% indicating complete user satisfaction with the multimedia data, and showed that the dependency between human receptivity and the required bandwidth of multimedia clips is non-linear [APT95]. In the context of bandwidth-constrained environments, results suggest that a limited reduction in human receptivity facilitates a relatively large reduction in bandwidth requirement (the *asymptotic property* of the VCS curves [GHP00]). This asymptotic property is clearly visible in Figure 2.12.

Ghinea and Thomas (1998): To measure the impact of video Quality of Service (QoS) variation on user perception and understanding of multimedia video clips, Ghinea and Thomas [GHI98] presented users with a series of 12 windowed (352*288 pixel) MPEG-1 video clips, each between 26 and 45 seconds long, with a consistent objective sound quality. The clips were chosen to cover a broad spectrum of subject matter, whilst also considering the dynamic, audio, video and textual content of the video clip. They varied the frame per second (fps) QoS parameters, whilst maintaining a constant colour depth, window size and audio stream quality. Frame rates of 25 fps, 15 fps and 5 fps were used and were varied across the experiment, yet for a specific user they remained constant throughout. 10 users were tested for each frame rate. Users were kept unaware of the frame rate being displayed. To allow dynamic (D), audio (A), video (V) and textual (T) considerations to be taken into account, in both questionnaire design and data analysis, characteristic weightings were used on a scale of 0-2, assigning importance of the inherent characteristics of each video clip. Table 2.3 contains the characteristic weightings, as defined by Ghinea and Thomas [GHI98].

Table 2.3: Video characteristics defined by Ghinea and Thomas [GHI98].

Video Category	Dynamic (D)	Audio (A)	Video (V)	Text (T)
1 – Commercial	1	2	2	1
2 – Band	1	2	1	0
3 – Chorus	0	2	1	0
4 – Animation	1	1	2	0
5 – Weather	0	2	2	2
6 – Documentary	1	2	2	0
7 – Pop Music	1	2	2	2
8 – News	0	2	2	1
9 – Cooking	0	2	2	0
10 – Rugby	2	1	2	1
11 – Snooker	0	1	1	2
12 – Action	2	1	2	0

The clips were chosen to present the majority of individuals with no peak in personal interest, which could skew results. Clips were also chosen to limit the number of individuals watching the clip with previous knowledge and experience. After the user had been shown a video clip, the video window was closed, and questions were asked about the video that they had just seen. The number of questions was dependent on the video clip being shown and varied between 10 and 12. Once a user had answered all questions relating to the video clip, and all responses had been noted, users were asked to rate the quality of the clip using a 6 point Likert scale, with scores 1 and 6 representing the worst and, respectively, the best possible perceived level of quality. Users were instructed not to let personal bias towards the subject matter influence their quality rating of the clip. Instead they were asked to judge a clip's quality by the degree to which they, the users, felt satisfied with the service of quality. The questions, used by Ghinea and Thomas, were designed to encompass all aspects of the information being presented in the clips: *D, A, V, T.* A number of questions were used to analyse a user's ability to absorb multiple media at one point in time; as correct answers could only be given if a user had assimilated information from multiple media. Lastly, a number of the questions were used that couldn't be answered by observation of the video alone, but by the users making inference and deductions from the information that had just been presented.

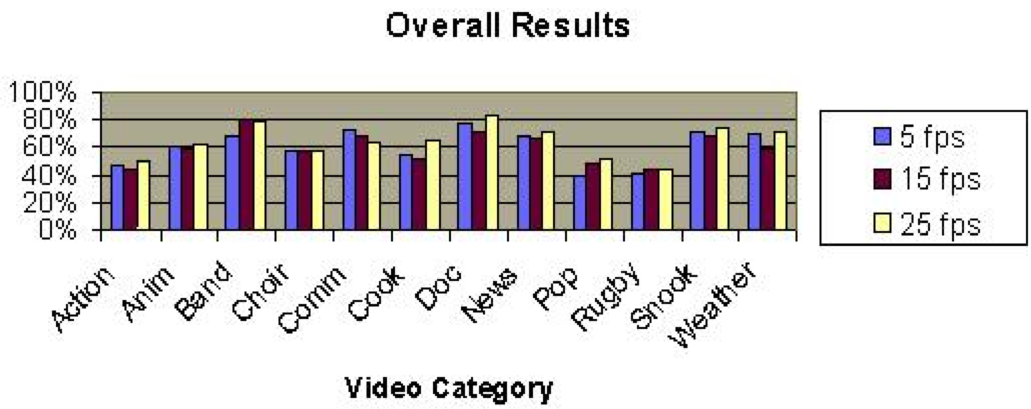


Figure 2.13: Effect of Varied QoS on User QoP [GHI98].

The main conclusions of this work include the following:

- A significant loss of frames (that is, a reduction in the frame rate) does not proportionally reduce the user's understanding and perception of the presentation (see Figure 2.13). In fact, in some instances the participant seemed to assimilate more information, thereby resulting in more correct answers to questions. Ghinea and Thomas proposed that this was because the user has more time to view a specific frame before the frame changes (at 25 fps, a frame is visible for only 0.04 sec, whereas at 5 fps a frame is visible for 0.2 sec), hence absorbing more information.

- Users have difficulty in absorbing audio, visual and textual information concurrently. Users tend to focus on one of these media at any one moment, although they may switch between the different media. This implies that critical and important messages in a multimedia presentation should be delivered in only one type of medium.

- When the cause of the annoyance is visible (such as lip synchronisation), users will disregard it and focus on the audio information if considered contextually important.

- Highly dynamic scenes, although expensive in resources, have a negative impact on user understanding and information assimilation. Questions in this category obtained the least number of correct answers. However the entertainment values of such presentations seem to be consistent, irrespective of the frame rate at which they are shown. The link between entertainment and content understanding is therefore not direct.

Ghinea and Thomas's method of measuring user perception of multimedia quality, later termed Quality of Perception (QoP), incorporates both a user's capability to understand the informational content of a multimedia video presentation, as well as his/her satisfaction with the quality of the visualised multimedia. QoP has been developed at all quality abstractions of our model (network-, media- and content-level) in cooperation with a number of authors: Fish [GHT00], Gulliver [GULL03; GULL04], Magoulas [GHM01] and Thomas [GHI99; GHI00; GHT00; GHT01]; concerning issues including the development of network protocol stacks, multimedia media assessment, attention tracking, as well user accessibility.

Gulliver and Ghinea (2003): [GULL03] used an adapted version of QoP to investigate the impact that hearing level has on user perception of multimedia, with and without captions. They showed users a series of 10 windowed (352*288 pixel) MPEG-1 video clips, each between 26 and 45 seconds long, with a consistent sound quality and video frame rate. Additional captions were added in a separate window, as defined by the experimental design.

Results showed that deafness significantly impacts a user's ability to assimilate information (see Figure 2.14). Interestingly, use of captions does not increase deaf information assimilation, yet increases quality of context-dependent information assimilated from the caption / audio.

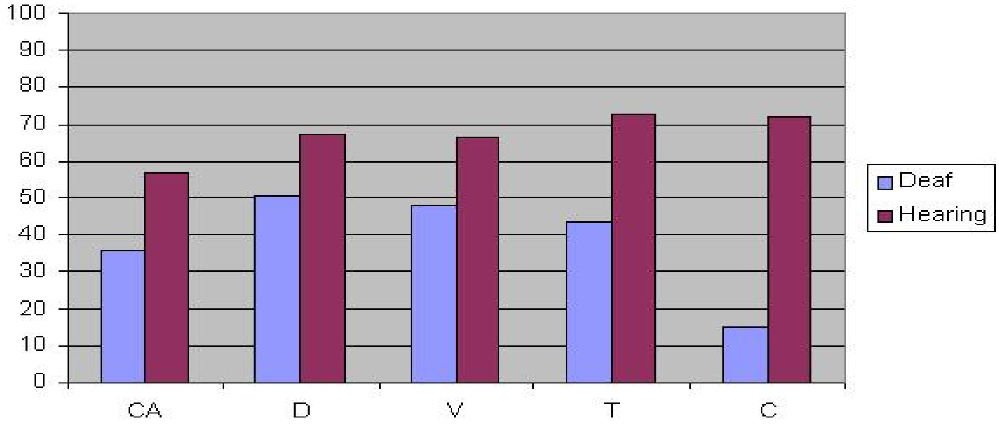


Figure 2.14: A detailed breakdown of Deaf / Hearing Information Assimilation (%): (CA) caption window / audio, (D) dynamic information, (V) video information, (T) textual information and (C) captions contained in the video window [GULL03].

To measure satisfaction Gulliver and Ghinea [GULL03] used two 11-point scales (0-10) to measure Level of Enjoyment (QoP-LoE) and user self-predicted level of Information Assimilation (QoP-PIA). A positive correlation was identified between QoP-LoE and QoP-PIA, independent of hearing level or hearing type, showing that a user's perception concerning there ability to assimilate information is linked to his/her subjective assessment of enjoyment.

Hollier and Voelcker (1997): Hollier and Voelcker [HOLL97] presented video clips supported by an audio commentary, and showed that participant quality rating with reference to audio commentaries is dependent on the quality of the video clips. Also, Hollier and Voelcker identified that the audio quality was found to have an influence on the perceived quality of the video presentation. Moreover, user perception of multimedia quality was found to be dependent on task. If users were involved in a learning task, then results suggest that audio quality needs to be substantially higher than when performing a routine or repetitive task [WAS96].

Video quality is considered as being more important in an interview situation, where unspoken communication may be of significant importance. However, perception of audio and video quality appears to be directly linked to the level of quality assumed to be necessary for a specific situation, suggesting that users have predefined quality criteria.

Kawalek (1995): Kawalek [KAW95] showed that loss of audio information has a more noticeable effect on the assimilation of informational content than video frame loss [KAW95]. Users are therefore less likely to notice degradation of video clips if shown low quality audio media (see Figure 2.15).

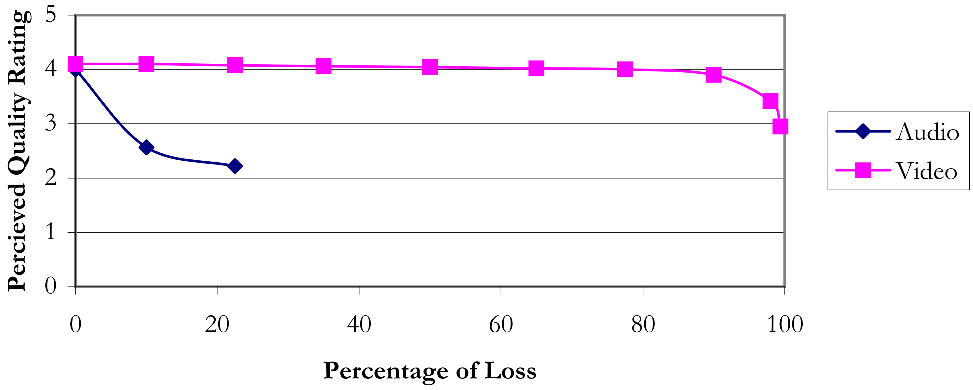


Figure 2.15: Perceived Effect of Transmission Quality on User Perception [KAW95].

Kies et al. (1997): Kies et al. [KIE97] conducted a two-part study to investigate the technical parameters affecting a *Desktop Video Conferencing system* (DVC). Consequently, three frame rate conditions (1, 6 and 30 fps), two resolution conditions (160x120 and 320x240), and three-communication channel conditions were manipulated. Dependent measures included the results of a questionnaire and subjective satisfaction, specifically concerning the video quality. Like Ghinea and Thomas [GHI98] and Procter et al. [PRO99], results suggest that factual information assimilation does not suffer under reduced video QoS, but subjective satisfaction is significantly decreased. In addition, a field study was used to look at the suitability of DVC for distance learning. Interestingly, field studies employing similar dependent measures indicated that participants may be more critical of poor video quality in laboratory settings.

Steinmetz (1996): Distributed multimedia synchronisation comprises both the definition and the establishment of temporal relationships amongst media types. In a multimedia context this definition can be extended such that synchronisation in multimedia systems comprises content, spatial and temporal relations between media objects. Perceptually, synchronisation of video and textual information or video and image information can be considered as either: *overlay*, which is information that is used in addition to the video information; or *no overlay*, which is information displayed, possibly in another box, to support the current video information. Blakowski and Steinmetz distinguished two different types of such media objects [BLA96]:

- *Time-dependent media objects*: these are media streams that are characterised by a temporal relation between consecutive component units. If the presentation duration of all components of an object is equal, then it is called a *continuous media object.*

- *Time-independent media objects*: these consist of media such as text and images. Here the semantics of their content does not depend on a time-structures.

An example of synchronisation between continuous media would be the synchronisation between the audio and video streams in a multimedia video clip. Multimedia synchronisation can, however, comprise temporal relations between both time-dependent and time-independent media objects. An example of this is a slide presentation show, where the presentation of the slides has to be synchronised with the appropriate units of the audio stream. Previous work on multimedia synchronisation was done in [BLA92; STE92], as well as on topics devoted to device synchronisation requirements [STE90]. In this study Steinmetz primarily manipulated media skews to measure how lip pointer and non-synchronisation impact user perception of what is deemed 'out of synch' [STE96].

A presentation was considered as being "in synch" when no error was identified i.e. a natural impression. A presentation was considered as being "out of synch" when it is perceived as being artificial, strange or even annoying. Video material was based around a simulated newsreader and facilitated three different views: head, shoulder and body, which related to the relative proportion of the newsreader shown in the video window.

Steinmetz artificially introduced skew in interval steps of 40ms, i.e.: -120ms, -80ms, -40ms, 0ms, +40, +80, +120 (where minus represents audio behind video and plus represents audio ahead of video). 107 participants viewed 3 sessions from 3 different views, lasting approximately 45 minutes in total. The same questions were used for all tests. Four separate regions were identified:

- The 'in sync' region spans a skew between –80ms (audio behind video) and +80ms (audio ahead of video). In this region most candidates did not detect synchronisation error. Lip synchronisation can only be tolerated within these limits.

- The 'out of sync' region spans beyond a skew of –160ms and +160ms. Nearly everybody detected these errors.

- The response inside first 'transient' area, where audio is ahead of video, is affected by the view; with the closer view (head) proving easier to detect errors, particularly around the eyes and mouth.

- The response inside the second 'transient' area, where video is ahead of audio, is also affected by the view. In this region video ahead of audio could be tolerated more than audio ahead of video.

Further experiments incorporating variation in video content, as well as the use of other languages (Spanish, Italian, French and Swedish) showed no impact on results. Interestingly, Steinmetz did measure variation as a result of the participant group, implying that user experience of manipulating video affects a user's aptitude when noticing multimedia synchronisation errors.

To investigate pointer synchronisation, in Computer Supported Cooperative Work (CSCW) environments, Steinmetz used two-business reports that contained accompanying graphics. All participants had a separate graphics viewing window, however a shared pointer was used to aid discussion.

Pointer synchronisation is different from lip synchronisation as it is more difficult to detect. Whilst participants notice lip synchronisation skews between 40 and 160ms, noticeable pointer skew values lie between 250 ms and 1500 ms.

- The 'in-sync' region lies between (audio ahead of pointing) –750ms and (pointing ahead of audio) +500ms.
- The 'out of sync' are spans beyond –1000ms and beyond +1250ms. From the user-perspective, greater skews are not deemed as being acceptable.
- Within the 'transient' areas candidates notice the 'out of sync' effect, but it was not mentioned as annoying.

Rimell and Hollier (1999): Rimell and Hollier [RIM99] examined the degree of interaction between auditory and visual human senses. Consequently, 72 participants were asked to grade video quality in a passive viewing situation. Experiments were divided into three sections according to the content of the video: Animation, Talking Head (such as a news reader) and general TV footage. Participant MOS was used as a reference to user perception of multimedia quality. The main findings of this study can be summarised as follows:

- Cross-modal interaction is present for all video, yet is particularly prolific when shown the talking head (close-up human) video.
- Steinmetz showed that humans are particularly sensitive to the distortion in human faces – particularly around the eyes and mouth [STE96]. Rimell and

Hollier concluded that minimising distortion in regions of the image around the eyes and mouth would help improve perceived quality for any given bandwidth environment.

- The relationship between perceived audio / video quality is dependent on video content, which highlights the need for perceptual measurement at the content-level quality abstraction.

Masry et al. (2001): Masry et al. [MAS01] performed a subjective quality evaluation to quantify viewer defects that appeared in low bit rate video at full and reduced frame rates. Eight thirty-second video sequences were used, which had been selected from various feature films and news/sport highlight collections, to cover a range of video content. Videos were compressed using three motion compensated encoders that operated at five bit/frame rate combinations - the majority of encoded sequences exhibited obvious coding artefacts. The subjective evaluation was performed using the SSCQE method. Viewers watched concatenated coded test sequences whilst continuously registering their perceived quality using a slider device. Results showed that both blockiness and blurriness resulted in consistent negative user quality perception. The affect of changes in frame rate on perceived quality are found to be related to the nature of the motion in the sequence, i.e. sequences with jerky motion benefited from the increased spatial quality at lower frame rates. The perceived qualities of sequences with smoother motion were in general unaffected by changes in frame rate.

Wijesekera et al. (1999): A number of mathematical measures of QoS (Quality of Service) models have been proposed [TOW93; WIJ96]. Wijesekera et al. [WIJ99] investigated the perceptual tolerance to discontinuity caused by media losses and repetition. Moreover, Wijesekera et al. considered the perceptual impact that varying degrees of synchronisation error have across different streams. Wijesekera et al., following the methodology of Steinmetz [STE96] - that is the manipulation of media skews, to measure stream continuity and synchronisation in the presence of media losses [WIJ96] - and consequently, quantifies human tolerance of transient continuity and synchronisation losses with respect to audio and video media. Wijesekera et al. found that:

- Viewer discontent with aggregate losses (i.e. the net loss, over a defined duration) gradually increases with the amount of loss, as long as losses are evenly distributed. For other types of loss and synchronisation error, there is a sharp initially rise in user discontent (to a certain value of the defect), after which the level of discontent plateaus.

- When video is shown at 30fps, an average aggregate loss below 17% is imperceptible, between 17% and 23% it is considered tolerated, and above 23% it is considered unacceptable, assuming losses are evenly distributed.

- Loosing two or more consecutive video frames is noticed by most users, when video is shown at 30fps. Loosing greater than two consecutive video frames does not proportionally impact user perception of video, as a quality rating plateau is reached. Similarly, loss of three or more consecutive audio frames was noticed by most users. Additional consecutive loss of audio frames does not proportionally impact user perception of audio, as a quality rating plateau is reached.

- Humans are not sensitive to video rate variations. Alternatively, humans have a high degree of sensitivity to audio, thus supporting the findings of Kawalek [KAW95]. Wijesekera et al. suggest that even a 20% rate variation in a newscast-type video does not result in significant user dissatisfaction. Where as a 5% rate variation in audio is noticed by most observers.

- Momentary rate variation in the audio stream, although initially considered as being amusing, was soon considered as annoying. This resulted in participants concentrating more on the audio defect than the audio content.

- An aggregated audio-video synchronisation loss of more that 20% frames was identified. Interestingly, consecutive synchronisation loss of more than 3 frames is identified by most users, which is consistent with [STE96].

Wilson and Sasse (2000): Bouch et al. [BOU01] proposed a 3-dimensional approach to assessment of audio and video quality in networked multimedia applications: measuring task performance, user satisfaction and user cost (in terms of physiological impact). Bouch et al. used their approach to provide an integrated framework from which to conduct valid assessment of perceived QoS (Quality of Service). Wilson and Sasse [WIL00b] used this 3-dimensional approach and measured: Blood Volume Pulse (BVP), Heart Rate (HR) and Galvanic Skin Resistance (GSR), to measure the stress caused when inadequate media quality is presented to a participant. Twenty-four participants watched two recorded interviews conducted, using IP video tools, lasting fifteen minutes each. After every five minutes the quality of the video was changed, allowing quality variation over time. Audio quality was not varied. Participants therefore saw two interviews with video frame rates of 5-25-5 fps and 25-5-25 fps respectively. Whilst viewing the videos, participants rated the audio / video quality using the QUASS tool, a *SSCQE* system where the participant continuously rated quality on an unlabelled scale. Physiological data was taken throughout the experiment. Moreover, to measure whether users perceived any changes in video quality, a questionnaire was also included. Wilson and Sasse showed that the GSR, HR and BVP data represented significant increases in stress when a video is shown at 5fps in comparison to 25fps. Only 16% of participants noticed a change in frame rate. No correlation was found between stress level and user feedback of perceived quality.

Subsequently, Wilson and Sasse [WIL00a] showed that subjective and physiological results do not always correlate with each other, which indicates that users cannot consciously evaluate the stress that degraded media quality has placed upon them.

Wikstrand and Eriksson (2002): Wikstrand and Eriksson [WIK02] used animation to identify how alternative rendering techniques impact user perception and acceptance, especially in bandwidth constrained environments. An animation or model of essential video activity demands that only context dependent data is transferred to the user and therefore reduces data transfer. Wikstrand and Eriksson performed an experiment to contrast different animations and video coding in terms of their cognitive and emotional effectiveness when viewing a football game on a mobile phone. Results showed that different rendering of the same video content affects the user's understanding and enjoyment of the football match. Participants who preferred video to animations did so because it gave them a better "football feeling", while those who preferred animations had a lower level of football knowledge and thought that animations were best for understanding the game. Wikstrand and Eriksson concluded that more advanced rendering, at the client end, may be used to optimise or blend between emotional and cognitive effectiveness.

Winkler (2001): Winkler [WIN01] incorporated visually appealing attributes (sharpness and colourfulness) in a video quality metric in order to broaden the factors used to determine the overall perceived quality of video. The results of subjective experiments showed that combining predictions with sharpness and colourfulness ratings leads to an improvement in automated quality prediction.

Media-Level Quality Variation (User-Perspective)

The media-level is concerned with how the media is coded for the transport of information over the network and / or whether the user perceives the video as being of good or bad quality. Accordingly, quality related studies adapting media, as a direct result of the user, are limited. The best example of quality related user media variation concerns attentive displays, which manipulate video quality around a user's point of gaze. Attentive displays offer considerable potential for the reduction of network resources and facilitate media-level quality variation with respect of both video content (technical-perspective) and user feedback (user-perspective) data. Further literature concerning attentive displays will be considered in chapter 5.

Media-Level Quality Measurement (Technical-Perspective)

- Ardito et al. [ARD94] developed a metric, which aimed to produce a linear mathematical model between technical and user subjective assessment.

- Watson [WAT98] and Xiao [XIA00] developed the *Digital Video Quality (DVQ)* Metric and *Video Quality Metric (VQM)* respectively. Both use Discrete Cosine Transforms (DCTs) to calculate the distance between two video clips (pre/post adaptation), and calculate a quality rating based on *Human Visual System* (HVS) attributes. Xiao also studied how compression options (quantisation parameter, quantisation matrix, spatial, scalability and frame drop) affect the video quality [XIA00].

- Winker [WIN01] and Quaglia and De Martin [QUA02] use *Peak Signal-to-Noise Ratio* (PSNR) as an objective measure of quality. The PSNR, however, does not consider user perception.

- Teo and Heeger [TE094] developed a normalised model of human vision. Extensions to the Teo and Heegar model included: the *Colour Masked Signal to Noise Ratio* metric (CMPSNR - Van den Branden Lambrecht and Farell, [VAN96]), used for measuring the quality of still colour pictures, and the *Normalised Video Fidelity Metric* (NVFM – Lindh and Van den Branden Lambrecht [LIN96]), for use with multimedia video. Extensions of the CMPSNR metric include: the *Moving Picture Activity Metric* (MPAM – Verscheure and Hubaux [VEH96]), the *Perceptual Visibility Predictor* metric (PVP - Verscheure, van den Branden Lambrecht [VEV97]) and the *Moving Pictures Quality Metric* (MPQM - van den Branden Lambrecht and Verscheure [VAV96]).

- Wang, Claypool and Zuo [WAN01] used frame rate as an objective measure of quality. Although the impact of frame rate is adapted by [APT95; GHI98; KAW95; KIE97; MAS01; WIJ99], Wang et al. is, to the best of our knowledge, the only study that used output frame rate as the quality criterion.

Ardito et al. (1994): The RAI Italian Television metric attempts to form a linear objective model from data representing subjective assessments concerning the quality of compressed images [ARD94; GHP00]. During subjective assessment the participants were presented with a sequence of pairs of video clips, one representing the original image and the other showing the degraded (compressed) equivalent. The user is not told which of the two is the original, yet is asked to categorise the quality of the two images using a five point Likert double stimulus impairment scale classification similar to the CCIR Rec. 500-3 scale [CCI74], with scores of 1 and 5 representing the "very annoying" and, respectively, "imperceptible" difference between the original and degraded images. All results are then normalised with respect to the original.

The RAI Italian Television metric initially calculates the SNR for all frames of the original and degraded video clips. To enable processing over time (the temporal variable) the SNR values are calculated across all frames, as well as in subsets of specified length *l*. Minimum and maximum values of the SNR are then determined for all groups with *I* frames, thus highlighting noisy sections of video. RAI Italian Television metric considers human visual sensitivity, by making use of a Sobel operator. A Sobel operator uses the luminance signal of surrounding pixels, in a 3*3 matrix, to calculate the gradient in a given direction. Applied both vertically and horizontally, Sobel operators can identify whether or not a specific pixel is part of an edge. The 3x3 matrix returns a level of luminance variation greater or smaller, respectively, than a defined threshold. An 'edge image' for a particular frame can be obtained by assigning a logical value of 1 or 0 to each pixel, depending on its value relative to the threshold. An edge image is defined for each frame, facilitating the calculation of SNR for three different scenarios: across the whole frame, only for areas of a frame belonging to edges, and, finally, across the whole frame but only for those areas not belonging to edges. Results show that, the RAI Italian Television linear model can successfully capture 90% of the given subjective information. However, large errors occur if the subjective data is applied across multiple video clips [ARD94], implying high content dependency.

Quaglia and De Martin (2002): Quaglia and De Martin [QUA02] describe a technique for delivering 'nearly constant' perceptual QoS when transmitting video sequences over IP Networks. On a frame-by-frame basis, allocation of premium packets (those with a higher QoS priority) depends upon on the perceptual importance of each MPEG *macroblock*, the desired level of QoS, and the instantaneous network state. Quaglia and De Martin report to have delivered nearly constant QoS, however constant reliance on PSNR and use of frame-by-frame analysis raises issues when considering the user perception of multimedia quality.

Teo and Heeger (1994): Teo and Heeger [TEO94] present a perceptual distortion measure that predicts image integrity based on a model of the human visual system that fits empirical measurements of: i) the response properties of neurons in the primary visual cortex, and 2) the psychophysics of spatial pattern detection, that is a person's ability to detect a low contrast visual stimuli.

The Teo Heeger model consists of four steps:

- *a front-end hexagonally sampled quadrature mirror filter transform* function [SIM90] that provides an output similar to that of retina, and is similarly tuned to different spatial orientations and frequencies.

- *squaring* to maximise variation.
- *a divisive contrast normalisation mechanism*, to represent the response of a hypothetical neuron in the primary visual cortex.
- *a detection mechanism* (both linear and non-linear) to identify differences (errors) between the encoded image and the original image.

Participants rated images, coded using the Teo and Heeger perceptual distortion measure, as being of considerably better 'quality' than images coded with no consideration to the user-perspective. Interestingly, both sets of test images containing similar RMS and PSNR values.

van den Branden Lambrecht and Farrell (1996): van den Branden Lambrecht and Farrell [VAF96] introduce a computation metric, which is termed the Colour Masked Signal to Noise Ratio (CMSNR). CMSNR incorporates opponent-colour (i.e. the stimulus of P-ganglion cells), as well as other aspects of human vision involved in spatial vision including: perceptual decomposition (Gabor Filters), masking (by adding weightings to screen areas), as well as the weighted grouping of neuron outputs. van den Branden Lambrecht and Farrell subsequently validated the CMPSNR metric, using 400 separate images, thus proving the CMPSNR metric as more able to predict user fidelity with a level of accuracy greater than the mean square error.

van den Branden Lambrecht and Verscheure (1996): van den Branden Lambrecht and Verscheure [VAV96] present the Moving Picture Quality Metric (MPQM) metric to address the problem of quality estimation of digital coded video sequences. The MPQM metric is based on a multi-channel model of the human spatio-temporal vision with parameters defined through interpretation of psychophysical experimental data. A spatio-temporal filter bank simulates the visual mechanism, which perceptually decomposes video into phenomena such as contrast sensitivity and masking. Perceptual components are then combined to produce a quality rating, by applying a greater summation weighting for areas of higher distortion. The quality rating is then normalised (on a scale from 1 to 5), using a normalised conversion [COM90]. van den Branden Lambrecht and Verscheure showed MPQM (moving picture quality metric) to model subjective user feedback concerning coded video quality.

Lindh and van den Branden Lambrecht (1996): Lindh and van den Branden Lambrecht [LIN96] introduced the NVFM model (Normalization Video Fidelity Metric), as an extension of the normalization model used by Teo and Heeger. The NVFM output accounts for normalization of the receptive field responses and inter-channel masking and is mapped onto the 1 to 5 quality scale on the basis of the vision model used in the MPQM metric.

Lindh and van den Branden Lambrecht compared NVFM with the Moving Picture Quality Metric (MPQM). Interestingly, the results of NVFM model are significantly different from the output of the MPQM, as it has a fast increase in user perceived quality in the lower range of bit rate, i.e. a slight increase of bandwidth can result in a very significant increase in quality. Interestingly, saturation occurs at roughly the same bit rate for both metrics (approximately 8 Mbit/sec). Lindh and van den Branden Lambrecht proposed NVFM as a better model of the cortical cell responses, compared to the MPQM metric.

Verscheure and van den Branden Lambrecht (1997): Verscheure and van den Branden Lambrecht [VEV97] considered the optimisation of video services by improving the bit allocation in a video encoder. This work is a continuation of the approach proposed in [VEB96], where Verscheure et al. introduced a local video activity metric (MPAM – Moving Picture Activity Metric), which accounts for both spatial and temporal perceptual activities, and can be used independently of the video encoding process. Verscheure and van den Branden Lambrecht introduce a more efficient block-based video metric, which is termed the perceptual visibility predictor (PVP). The PVP aims to classify local areas in terms of their relevance to human perception by incorporating MPQM (Moving Pictures Quality Metric). Simulations showed that PVP permits a 9% reduction in bandwidth, at the MPEG-2 bit allocation stage, without negatively affecting user perception of multimedia quality.

Watson (1998): Watson [WAT98] developed a number of visual quality metrics for evaluating, controlling and optimising the quality of compressed still images [WAT94; WAT95; WAT97; WAY97]. These metrics incorporate a simplified model of the human visual sensitivity to spatial and chromatic visual signals. Watson extended his work to propose a new video quality metric, called the *Digital Video Quality (DVQ)* metric. The DVQ metric is based on the Discrete Cosine Transform (DCT) and computes the visibility of artefacts in the DCT domain using a multistage analysis of video frames. Watson purposefully minimized the amount of memory and computation effort required by the metric, in order that it might be applied to a range of applications.

Xiao (2000): Xiao [XIA00] developed a modified DCT-based *video quality metric* (VQM), based on Watson's DVQ metric. Instead of applying temporal filtering and human spatial contrast sensitivity function separately, Xiao considering static sensitivity and dynamic interaction of frames in one step. Xiao shows that VQM identifies artefacts that are not identified by RMS.

Media-Level Quality Measurement (User-Perspective)

- Apteker et al. [APT95], measured 'watchability' (receptivity) as a measure of user satisfaction concerning video quality.

- Ghinea and Thomas [GHI98], asked respondents to rate the quality of each clip on a seven point Likert scale.

- Procter et al. [PRO99] ask subjects to compare the streamed video against the non-degraded original video. Quality was measured by asking participants to consider a number of statements and, using a seven-point Likert-style scale, e.g. "the video was just as good as watching a live lecture in the same room" and "the video was just as good as watching a VCR tape on a normal television".

- Wilson and Sasse [WIL00a; WIL00b] used the *Single Stimulus Continuous Quality* QUASS tool to allow the user to continuously rate the audio / video quality, whilst viewing a video presentation.

- Wikstrand and Eriksson [WIK02] measured user preference concerning the animation rendering technique.

- Winkler [WIN01] simultaneously showed participants the original and degraded video clips, using the *Double Stimuli Continuous Quality Scale* (DSCQS) method, to find out which participants preferred. The quality of each video was rated on an unmarked scale from "bad" to "excellent". From the relative difference, a *Differential Mean Opinion Score* (DMOS) was calculated.

- Steinmetz [STE96] used participant annoyance of synchronisation skews as a measure of quality. In both cases only identified errors are considered as being of low quality.

2.3.3.4. The Content-Level

Content-Level Quality Variation (Technical-Perspective)

- Ghinea and Thomas [GHI98], Gulliver and Ghinea [GUL03], Masry et al. [MAS01] , Rimell and Hollier [RIM99], as well as Steinmetz [STE96] all varied experimental material to ensure diverse media content.

- Procter et al. [PRO99] used diametrically opposed presentations: a bank's annual report and a dramatized scene of sexual harassment.

- Steinmetz [STE96] used three different views: head, shoulder and body, which related to the relative proportion of the newsreader shown in the video window.

Content-Level Quality Variation (User-Perspective)

- Gulliver and Ghinea [GUL03] varied participant demographics to measure changes in multimedia perception as a result of deafness, deafness type and use of captions.

- Steinmetz [STE96] tested videos using a variety of languages (Spanish, Italian, French and Swedish) in order to check lip synchronisation errors.

- Watson and Sasse [WAS00] varied peripheral factors, such as volume and type of microphone, to measure in a CSCW environment, the impact on user perception of audio quality.

- Wikstrand and Eriksson [WIK02] adapted animation rendering techniques, whilst maintain important presentation content.

Watson and Sasse (2000): Watson and Sasse [WAS00] showed that volume discrepancies, poor quality microphones and echo have a greater impact on a user's perceived quality of network audio than packet loss.

Content-Level Quality Measurement (Technical-Perspective)

- Wilson and Sasse [WIL00a; WIL00b] measure participants Blood Volume Pulse (BVP), Heart Rate (HR) and Galvanic Skin Resistance (GSR), to measure for stress as a result of low quality video.

Content-Level Quality Measurement (User-Perspective)

- Apteker et al. [APT95], measured 'watchability' (receptivity) as a measure of user satisfaction concerning video content along temporal, visual and audio dimensions. Accordingly, 'watchability' covers both media- and content-levels.

- Procter et al. [PRO99] used 'ease of understanding', 'recall', 'level of interest', and 'level of comprehension' as quality measures.

- Ghinea and Thomas [GHI98], Gulliver and Ghinea [GUL03] used questionnaire feedback to measure a user's ability to assimilate and understand multimedia information.

- Gulliver and Ghinea [GUL03] asked participants to predict how much information they had assimilated during IA tasks, using scores of 0 to 10 representing "none" and, respectively, "all" of the information that was perceived as being available. Gulliver and Ghinea also measured a participant's level of enjoyment, using scores of 0 to 10 representing "none" and, respectively, "absolute" enjoyment.

- Wikstrand and Eriksson [WIK02] showed that animation rendering affects user's understanding and enjoyment of a football match.

2.4. Achieving a Extensive User-Perspective

Inclusion of the user-perspective is of paramount importance to the continued uptake and proliferation of multimedia applications since users will not use and pay for applications if they are perceived to be of low quality. Section 2.3 highlighted the diverse range of studies that have conducted to assess multimedia quality. Interestingly, no extensive set of studies have been undertaken that consistently measure the infotainment duality of distributed multimedia quality.

2.4.1. Reviewing the Literature

Section 2.3 has highlighted a number of studies that measure the user-perspective at the content-level (Apteker et al. [APT95], Ghinea and Gulliver [GUL03], Ghinea and Thomas [GHI98], Procter et al. [PRO99], Wilson and Sasse [WIL00a; WIL00b]). These are summarised in Table 2.4, which:

i) Lists the primary studies that measure the user-perspective at the content-level, stating the number of participants used in each study.

ii) Identifies the adapted quality parameters, and defines the quality abstraction at which each parameter was adapted (N = Network-level, M = Media-level, C = Content-level).

iii) Provides a list of the measurements taken for each study and the quality level abstraction at which each measurement was taken (N = Network-level, M = Media-level, C = Content-level).

Table 2.4: Comparison of User Perceptual Studies
[APT95; GHI98; GUL03; PRO99; WIL00a; WIL00b].

Study	Participants	Adapted	Measured
Aptker et al. [APT95]	60 students	• Frame rate (M) • Video Content (C)	• Watchability (M)(C)
Gulliver and Ghinea [GUL03]	50 participants (30 hearing / 20 deaf)	• FrameRate (M) • Captions (M) • Video Content (C) • Demographics (C)	• Information Asimilation (C) • Satisfaction (C) • Self perceived ability (C)
Procter et al. [PRO99]	24 participants	• Network Load (N) • Video Content (C)	• Comprehension (C) • Uptake of non-verbal information (C) • Satisfaction (M)
Wilson and Sasse [WIL00a; WIL00b]	24 participants	• Frame Rate (M)	• Galvanic Skin Resistance (C) • Heart Rate (C) • Blood Volume Pulse (C) • QUASS (M)
Ghinea and Thomas [GHI98]	30 participants	• Frame rate (M) • Video Content (C)	• Information Assimilation (C) • Satisfaction (M)

To extensively consider distributed multimedia quality effectively from a user-perspective it is essential that, where possible, both technical- and user-perspective parameter variation is made at all quality abstractions of our model, i.e. network-level (technical-perspective), media-level (technical- and user-perspective) and content-level (technical- and user-perspective) parameter variation – see Figure 2.11. Moreover, in order to effectively measure the infotainment duality of multimedia, i.e. information transfer and level of satisfaction, the user-perspective must consider both:

- the user's ability to assimilate / understand the informational content of the video {assessing the content-level user-perspective}.

- the user's satisfaction, both measuring the user's satisfaction with the objective QoS settings {assessing the media-level user-perspective}, and also user enjoyment {assessing the content-level user-perspective}.

Interestingly, none of the mentioned studies achieved this set of criteria and it is on this that our research shall focus its attention.

2.4.2. Research Aims and Objectives

We have identified a model of user quality consisting of three quality abstractions (network-level, the media-level and the content-level), viewed from two separate perspectives (the technical-perspective and user-perspective). Accordingly, we defined the following research aim: to extensively consider the user's perception of distributed multimedia quality, by adapting relevant technical- and user-perspective parameters at all quality abstractions: network-level (technical-perspective), media-level (technical- and user-perspectives), and content-level (technical- and user-perspectives). In addition, user perception of multimedia quality should consistently consider both how the multimedia presentation was assimilated / understood by the user at the content-level, yet also examine the user's satisfaction (both his/her satisfaction with the objective QoS settings {media-level} and enjoyment concerning the video content {content-level}).

To achieve the defined research aim, a series of three investigations, structured along the Network, Media and Content axes of our model, will be carried out, each targeting a major research objective of our study. Due to the reduced bandwidth requirement and increased perceptual impact of corrupted audio, in our research the audio stream will not be manipulated. By manipulating only video parameters we also minimise the number of variables that impact the users' perception of quality.

Objective 1: Measurement of the perceptual impact of network-level parameter variation. To this end, we intend to measure the impact of delay and jitter on user perception of multimedia quality. Additionally, eye-tracking technologies will be employed in our work to help identify how gaze disparity in eye-location is affected at the network-level. By continuously monitoring user focus we aim to gain a better understanding why people do not notice obvious cues in the experimental video material. Eye-tracking will be measured at the network-level, however, due to the complexity of eye-tracking data, will be analysed separately to QoP data. Although the impact of delay and jitter have been considered by other authors, these studies fail to considering both level of user understanding (information assimilation) and user satisfaction (both with the video QoS {media-level} and level of enjoyment concerning the content of the video {content-level}). Our work in this respect will be detailed in Chapter 4.

Objective 2: Measurement of the perceptual impact of media-level parameter variation. The *Human Visual System* (HVS) can only process detailed information within a small area at the centre of vision, with rapid acuity drop-off in peripheral areas [MAC70]. Attentive displays monitor and/or predict user gaze, to manipulate allocation of bandwidth, such that quality is improved around the point of gaze [BAR96]. Attentive displays offer considerable potential for the reduction of network resources and facilitates media-level quality variation that incorporates both video content-based (technical-perspective) and user-based (user-perspective) data. Accordingly, to consider media-level parameter variation, with consideration of both technical- and user-perspectives, we intend to measure the impact of using a region-of-interest attentive display system, which varies the frame-rate of certain regions of interest, as defined by both video content-based and user-based data. Eye-tracking data will be used at the media-level to manipulate video content, yet no monitoring of user eye-gaze location will be

made at the media-level. Our work in this respect will be detailed in Chapter 5.

Objective 3: Measurement of the perceptual impact of content-level parameter variation. To consider content-level user-perspective parameter variation we intend to measure the impact of display-type on user perception of multimedia quality. The perceptual impact of display-type has particular importance in the areas of virtual reality and mobile computing. To ensure technical-perspective variation at the content-level, diverse experimental material is used, and Chapter 6 describes this last component of our study. Although an eye-tracker will be used as a display device, no monitoring of user eye-gaze location data will be made at the content-level.

2.5. Conclusion

In this chapter we have introduced the reader to issues concerning perceptual studies of multimedia quality. We have identified a distinctive model of multimedia quality definition and have accordingly identified a need for an extensive study measuring the user-perspective at all quality abstractions, as a result of appropriate technical- and user-perspective quality parameter variation. The methodology employed in our three investigations will now be detailed and justified in the following chapter.

CHAPTER 3

RESEARCH METHODOLOGY

3.1. Introduction

In our work, we wish to explore the human side of the multimedia experience – the user-perspective. We therefore define the user-perspective as being not only the user's understanding concerning the multimedia presentation (content-level), but also his/her satisfaction concerning both the objective quality of the video (media-level) and the level of enjoyment (content-level). The purpose of this chapter is to describe the research methodology, used in chapters 4, 5 and 6, to extensively assess user-perspective of multimedia quality when appropriate technical- and user-perspective parameter variation is made at network-, media- and content-levels respectively. In chapter 2 we presented the reader with a background concerning the definition and measurement of distributed multimedia quality at different quality abstractions. Specifically, we have identified the following research aim.

> ***In this study, we aim to extensively consider the user's perception of distributed multimedia quality, by adapting relevant technical- and user-dependent parameters at all quality abstractions: network-level (technical-perspective), media-level (technical- and user-perspectives), and content-level (technical- and user-perspectives). In addition, user perception of multimedia quality should consistently consider both how the multimedia presentation was assimilated / understood by the user at the content-level, yet also examine the user's satisfaction (both his/her satisfaction with the objective QoS settings {media-level} and level of enjoyment concerning the video content {content-level}).***

This chapter begins by justifying the use of structured laboratory-based experiments to investigate the research aim. In section 3.3, the adapted experimental method is described and reasons are given for method variation, in order to achieve the defined research aim. In section 3.4, we then consider and justify the use of experimental material (multimedia videos). Finally, in section 3.5, we provide the user with information relating to eye tracking and its uses in our investigations.

3.2. Experimental Methodologies

3.2.1. Experiments: The obvious solution

Experimental methodology is used to determine significant differences between controlled conditions [COO94]. Accordingly, the experimental method is considered as most fitting to establish differences in user perception of quality as a result of parameter variation. The correct experimental design facilitates the removal of confounding variables, yet ensures an unparalleled level of consistency between studies.

User perception is realistically an undeterminable idea as no direct or complete measurement of user perception can be made. Use of an experimental method, however, provides a consistent measurement of certain aspects of user perception, which can be used to determine significant variation. Although the very act of measuring perception can indirectly influence user focus, and therefore perception itself, laboratory experiments ensure a level of consistency that is not possible using other methods [COO94].

3.2.2. Structured or Non-Structured Experimentation

This section determines the difference between subjective structured and non-structured experiments and considers the relevant benefits of both methods when used in context of our research aims and objectives.

Structured experiments are those that follow a predefined order or direction, with the ultimate aim of consistently measuring changes in pre-ordained experimental factors. Structured experiments require the design and use of questionnaires and/or checklists to extract relevant information to identify specific results or outcomes. Non-structured experiments do not necessarily require a clear-cut or pre-defined order or direction. Accordingly, non-structured experiments are neither consistent in content nor measure pre-designed experimental factors. Use of open interviewing or participant observation provides a rich source of information, partly as participant response is not placed within a rigid experimental structure, however as limited experimental organisation exists, results can only be analysed using interpretative methods. The choice between structured or non-structured experiments thus allows the choice between either rich (unstructured) or limited (structured) data.

As repeatability is central to ensuring a consistent user focus, and subsequently user perception, it is paramount to the successful implementation of our study that we rely primarily on structured experiments. We have identified a clear set of structured aims and objectives, which facilitates the effective measurement of critical experimental factors. Consequently, structured experimentation, incorporating predefined questionnaires, will be used throughout this study to determine the perceptual variation of participants as a result of quality parameter variation. In addition to experimental questionnaires, eye-tracking technology will be used, to enrich the understanding the impact of network-level parameter variation on user point of gaze. Eye-tracking is used in our work, as monitoring eye movements offers insights into user perception, as well as the associated attention mechanisms and cognitive processes. Interpretation of eye movement data can be based on the empirically validated assumption that when a person is performing a cognitive task, the location of his/her gaze corresponds to the symbol currently being processed in working memory [JUS76] and, moreover, that the eye naturally focuses on areas that are most likely to be informative [MAC70].

The use of eye-tracking in our study is motivated by the findings of Ghinea [GHP00], who showed that a user's assimilation of the informational content of clips is characterised by a WYS<>WYG (What You See is not equal to What You Get) relation. This means that users, whilst still absorbing some information correctly, do not notice obvious cues in the video clips. Instead users often appear to determine conclusions as a result of reasoning, arriving at their conclusions based on intuition and past experience. Ghinea [GHP00] proposed the use of eye-tracking in future network-level user-perspective studies to identify the reasons why people do not notice obvious cues in the experimental video material, thus providing a better understanding of the role that the human element plays in the reception, analysis and synthesis of multimedia data. Eye-tracking will be employed in our work to help identify how gaze disparity in eye-location is affected at the network-level. A small variation between the eye gaze locations of multiple users implies the existence of explicit visual cues. A great diversity in the location of eye gaze implies no obvious single point of focus. By continuously monitoring user focus we gain a better understanding why people do not notice obvious cues in the experimental video material. Eye-tracking will be measured at the network-level, yet due to the complexity of eye-tracking data will be analysed separately to QoP data. Eye-tracking data will be used at the media-level to manipulate video content, yet no monitoring of user eye-gaze location will be made at the media- and content-levels.

3.2.3. Laboratory or Field Studies

Using a structured experimental method facilitates the intentional manipulation of one or more independent variable(s). Consequently, the statistical effects of the independent variable(s) are measured in terms of change that occurs on one or more dependent variable(s). This section determines the difference between laboratory (true) and field experiments and considers the relevant benefits of both methods when used in the context of our aims and objectives.

Laboratory studies, also known as true experiments [COO94], control all variables that may confound or influence the current aims and objectives. To achieve this control over experimental variables, tasks and measures, an controlled environment is required. Although laboratory studies facilitate highly focused, consistent and accurate studies, Coolican [COO94] highlights four main weaknesses of the laboratory experiment:

i) the artificial environment.

ii) the reduced ability to generalise experimental results.

iii) the restricted list of variables.

iv) the reduced level of situational realism.

If, as part of a structured laboratory experiment, an unrealistic experimental design or an unnatural process is forced upon participants, then unreliable results are likely, e.g. if the user is asked to look out for particular items or perform a specific task then the focus of the user will be unnaturally affected [YAR67], which in turn affects the information assimilated by the user. Moreover, if experimental emphasis is placed on only particular variables, then use of structured laboratory experiments risks limiting the richness of feedback information. To prevent bias, Robson [ROB94] suggests minimal interaction between the participant and the experimenter.

In contrast, field experiments allow investigation within a real world environment. As with true experiments, field studies manipulate one or more independent variable(s) and measure the statistical effect of dependent variable(s). Although field studies can represent 'real-world' situations, they do not give the experimenter full control of the environment. Although confounding variables are almost impossible to control in a non-laboratory environment, field studies provide the possibility for a greater richness of data.

We have already stated that we plan to use eye-tracking data measurements at the network-level. Eye-tracking systems can be used as data gathering devices or can provide the user with interactive functionality [ISO00; REI02]. Depending on the equipment, they can be considered as either intrusive or non-intrusive in nature [GOL02] and can be developed as either pervasive [SOD02] or standalone systems. Accordingly, the eye-tracking device employed in our experiments significantly affects the use, implementation and data collection methodology. Due to cost and functional limitations, we were unable to use a mobile, head-mounted non-intrusive eye-tracking system, which would facilitate field experiments. Accordingly, realistic field experiments are not a practical possibility. To ensure the effective calibration and use of eye-tracking we require controlled laboratory experimentation. Furthermore, in order to minimise experimental bias, interaction between the experimenter and participant will be limited to the experimenter only asking the participant questions incorporated in the study questionnaire (see Appendix A).

3.3. Quality of Perception: An Adaptable Approach

3.3.1. Defining Quality of Perception

In order to explore the human side of the multimedia experience, we have used the *Quality of Perception* (QoP) concept. QoP is based on the idea that the technical-perspective alone is incapable of defining the perceived quality of multimedia video, especially at the content-level [BOU01; GHI98; WAS97]. In their research, Ghinea and Thomas [GHI98] presented users with a series of 12 windowed (352*288 pixel)

MPEG-1 video clips, each between 26 and 45 seconds long, with a consistent objective sound quality. The clips covered a broad spectrum of subject matter, with diversity defined along dynamic (D), audio (A), video (V) and textual (T) axis using a scale of 0-2, representing the importance of the inherent characteristics of each video clip (see Table 2.3). After the user had been shown a video clip, the video window was closed, and questions were asked about the video that they had just seen. The number of questions was dependent on the video clip being shown and varied between 10 and 12. Once a user had answered all questions relating to the video clip, and all responses had been noted, users were asked to rate the clip's perceived level of quality .

QoP is a concept that captures multimedia infotainment duality and more closely reflects multimedia's infotainment characteristics (i.e. that multimedia applications are located on the informational-entertainment spectrum). Quality of Perception uses user 'satisfaction' (QoP-S) and level of 'information assimilation' (QoP-IA) to determine the perceived level of multimedia quality. To this end, QoP encompass not only a user's satisfaction with the quality of multimedia presentations ('Satisfaction' - S), but also his/her ability to analyse, synthesise and assimilate the informational content of multimedia ('Information Assimilation' – IA).

As described above, Ghinea and Thomas originally defined QoP-S by measuring the user's perceived quality associated with the video presentation. Although determining the user perception of video quality at a media-level, variation to the original methodology was required, in order to measure user satisfaction at both media- and content-level quality abstractions. QoP was previously adapted by [GUL03], to measure the impact of hearing level of a user's level of enjoyment (QoP-LoE) and self-predicted level of information assimilation (QoP-PIA). Interestingly, as both QoP-LoE and QoP-PIA relate to the transfer of information to the user, they are measured at the content-level, which successfully demonstrates the ability of QoP-S to facilitate content-level user feedback.

In our study QoP-S is considered as being subjective in nature and consists of two component parts: QoP–LoQ (the user's judgement concerning the Level of Quality assigned to the multimedia content being visualised) and QoP–LoE (the user's Level of Enjoyment whilst viewing multimedia content), thus targeting quality perception at both the media- and content-levels respectively. Accordingly, QoP-S successfully considers the user-perspective from both user quality paradigms as defined in our research aim.

3.3.2. Measuring QoP

There are four basic methods to measure subjective perceptual quality [CLA99; GHI98; WIL98]: *Forced Choice* in which test subjects are presented the same clip under two different quality levels; *Content Query* in which test subjects are asked questions about the content of video clips after watching them; *Effectiveness Study* in which different quality videos are used for real life tasks and task effectiveness, then, becomes the measure of quality; and *Quality Opinion Score* in which test subjects are asked for an opinion score during or after watching a video clip. In our study we use content query and quality opinion scores to measure user QoP-IA (information assimilation and user QoP-S (satisfaction) respectively. To understand QoP it is important that the reader understands how QoP factors were defined and measured. These issues shall now be addressed.

3.3.2.1. Measuring User Information Assimilation / Understanding (QoP-IA)

QoP-IA implements content query and allows us to measure a user's ability to understand / assimilate the content of the video clip (content-level). QoP-IA was expressed as a percentage measure that reflects a user's level of understanding and information assimilation, from visualised multimedia content. Thus, after watching a particular multimedia clip, the user was asked a number of questions that examined the information being assimilated from certain information sources. For each feedback question, the source of the answer was determined as having been sourced by one or more of the following information sources:

V : Video-based information that comes from the video window, which does not contain text. Originally Ghinea and Thomas [GHI98] defined (V) and dynamic-based (D) information separately. However, as user feedback suggested that the distinction between these variables were confusing, these information sources were combined in our study.

A : Audio-based information that is presented in the audio stream.

T : Textual-based information that is contained in the video window, e.g. the newscaster's name in a caption window.

QoP-IA is calculated as being the percentage of correctly assimilated information. Consequently, all QoP-IA questions are designed so that specific information must be assimilated in order to correctly answer each question. The majority of questions may be assimilated from a single information source, however, a number of questions relate to multiple information sources. The following example, from the pop video clip (see Appendix B), shows how questions were used to test the user's assimilation and understanding of V, A and T information sources (the source of the data is contained in brackets and the answer is underlined):

- What was the bald man doing in the video? (V) Moving a chair / furniture.
- Name two features of the clip that relate to the Orient? (V) She is wearing a t-shirt that has a dragon logo, (T) She performed in a Japanese video commercial
- According to the lyrics of the song, is the male character on time? (A) He is late.
- What time was the clip broadcast? (T) 7:59 am

As questions have unambiguous answers, it is possible to establish whether a participant assimilated certain information from the video presentation. Accordingly, it is possible to calculate the percentage of correctly assimilated information, facilitating comparison between user information assimilation / understanding, as a result of quality parameter variation.

At this point, it may be argued that not all videos contain V, A and T information sources, which risks making QoP-IA a nonsensical measure. Moreover, the output of QoP-IA is dependent on the type of questions being used for each video clip, risking limitation of user feedback, because of a non-standard distribution of questions. Although these are important considerations, concerns are unfounded ensuring that: i) considerable variation in video material content; ii) consistency is ensured for all experiments; iii) a reasonable sample size is used to filter out individual differences; and iv) we focus on the statistical variation in user

QoP-IA as a result of independent experimental variable variation. QoP-IA is measured across a number of different video content, therefore as long as a consistent set of questions are used, and the sample size is adequately large, any significant variations in QoP-IA will still be as a result of a changes in the independent experimental variable(s).

3.3.2.2. Measuring Subjective User Satisfaction (QoP-S)

QoP-S is subjective in nature and consists of two component parts: QoP–LoQ (the user's judgement concerning the Level of Quality assigned to the multimedia content being visualised) and QoP–LoE (the user's Level of Enjoyment whilst viewing multimedia content).

To ensure that user satisfaction includes measurement at the media-level we have used QoP-LoQ (the user's judgement concerning the Level of Quality assigned to the multimedia content being visualised), the first component part of QoP-S in our approach. In order to measure QoP-LoQ, users were asked to indicate, on a scale of 0 - 5, how they judged, independent of the subject matter, the presentation quality of a particular piece of multimedia content they had just seen (with scores of 0 and 5 representing "no" and, respectively, "absolute" user satisfaction with the multimedia presentation quality). Accordingly, QoP-S incorporates media-level user-perspective, as defined in our research aim.

To ensure that user satisfaction includes measurement at the content-level we have used QoP-LoE (the subjective Level of Enjoyment experienced by a user when watching a multimedia presentation), which is the second and final component part of QoP-S in our study. To measure QoP-LoE, the user was asked to express, on a scale of 0 - 5, how much they enjoyed the video presentation (with scores of 0 and 5 representing "no" and, respectively, "absolute" user satisfaction with the multimedia video presentation). Accordingly, QoP-S incorporates the content-level user-perspective, as defined in our research aim.

0 – 5 single stimulus quality scale were used to assess QoP-S factors in order that results aligned with the International Telecommunications Union's appraisal of both audio (ITU-T recommendation [ITU80]) and video quality (ITU-R

recommendations [ITU50]). Continuous simultaneous assessment of QoP-S factors was considered, yet was deemed as being an unacceptable additional cognitive load.

3.4. Experimental Material

As well as fulfilling the need for content-level technical variation, the use of a range of video material is essential to ensure the extensive usability of QoP-IA. Accordingly, effective justification of experimental material is important to ensure that results can be appropriately generalised.

3.4.1. Original Material

In this section, we describe the set of video clips previously used by [GHI98; GUL03], who presented all participants with a series of 12 and 10 windowed MPEG video clips respectively. The duration of video clips used by [GHI98] was between 26 and 45 seconds long, which is ideal for use with QoP, as limitations of human working memory [ALD95] can cause users to forget information shown at the beginning of the clip, if duration is substantially longer than 30 seconds.

The multimedia video clips originally used with QoP were specifically chosen to cover a broad spectrum of infotainment. Moreover, the clips were chosen to present the majority of individuals with no peak in personal interest, whilst limiting the number of individuals watching the clip with previous knowledge and experience. The multimedia video clips used varied from those that are informational in nature (such as a news / weather broadcast) to ones those that are usually viewed purely for entertainment purposes (such as an action sequence, a cartoon, a music clip or a sports event). Specific clips were chosen as a mixture of the two viewing goals, such as the cooking clip. These videos were:

- **BA (Commercial clip)** - an advertisement for a bathroom cleaner is being presented. The qualities of the product are praised in four ways - by the narrator, both audio and visually by the couple being shown in the commercial, and textually, through a slogan display.

- **BD (Band clip)** - this shows a high school band playing a jazz tune against a background of multicoloured and changing lights.

- **CH (Chorus clip)** - this clip presents a chorus comprising 11 members performing mediaeval Latin music. A digital watermark bearing the name of the TV channel is subtly embedded in the image throughout the recording.

- **DA (Animation clip)** - this clip features a disagreement between two main characters. Although dynamically limited, there are several subtle nuances in the clip, for example: the correspondence between the stormy weather and the argument.

- **FC (Weather clip)** - this is a clip about forthcoming weather in Europe and the United Kingdom. This information is presented through the three main channels possible: visually (through the use of weather maps), textually (information regarding envisaged temperatures, visibility in foggy areas) and orally (by the presentation of the forecaster).

- **LN (Documentary clip)** - a feature on lions in India. Both audio and video streams are important, although there is no textual information present.

- **NA (Pop clip)** - is characterised by the unusual importance of the textual component, which details facts about the singer's life. From a visual viewpoint it is characterised by the fact that the clip was shot from a single camera position.

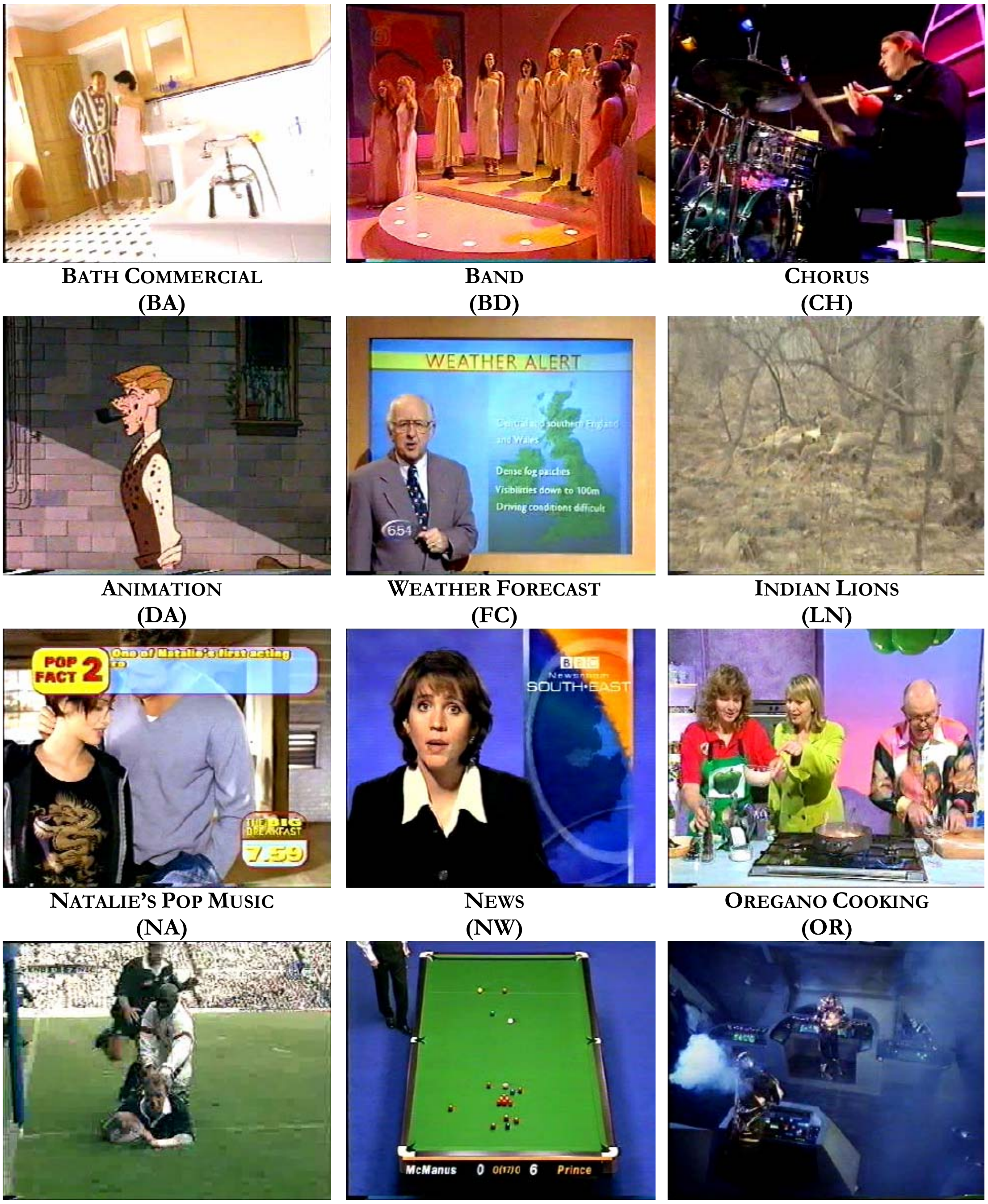
- **NW (News clip)** - contains two main stories. One of them is presented purely by verbal means, while the other has some supporting video footage. Rudimentary textual information (channel name, newscaster's name) is also displayed at various stages.

- **OR (Cooking clip)** - although largely static, there is a wealth of culinary information being passed to the viewer. This is done both through the dialogue being pursued and visually, through the presentation of ingredients being used in cooking of the meal.

- **RG (Rugby clip)** - presents a test match between England and New Zealand. Textual information (the score) is displayed in the upper left corner of the screen. The main event in the clip is the scoring of a try. The clip is characterised by great dynamism.

- **SN (Snooker clip)** - the lack of dynamism is in stark contrast to the Rugby clip. Textual information (the score and the names of the two players involved) is clearly displayed on the screen.

- **SP (Space clip)** - this was an action scene from a popular science fiction series. As is common in such sequences it involves rapid scene changes, with accompanying visual effects (explosions).



Bath Commercial (BA)

Band (BD)

Chorus (CH)

Animation (DA)

Weather Forecast (FC)

Indian Lions (LN)

Natalie's Pop Music (NA)

News (NW)

Oregano Cooking (OR)

Rugby (RG)

Snooker (SN)

Space (SP)

Figure 3.1: Shows video Frame 500, for the 12 video clips used in our experiment, demonstrating the diversity of multimedia being considered.

Figure 3.1 shows screen shots of the video clips used by Ghinea and Thomas [GHI98]; however, see Appendix B for more detailed information concerning frame images, dialogue content and dialogue timing.

3.4.2. Testing Experimental Material

The video material originally defined by [GHI98] (see Figure 3.1) should be used, if we intend to ensure consistency between QoP studies. However, as the content of video determines the use of questions, and ultimately the reliability of QoP-IA results, it is crucial that participants consider the video content of experimental material as significantly varied in content. Accordingly, a test was performed to identify user feedback in respect of video characteristic weightings defined by Ghinea and Thomas in Table 2.3.

Twenty-one participants, from a wide range of different ages (14 - 72) and backgrounds (students, blue and white collar workers, two nurses, some retired persons and even a reverend), were asked to watch the original video material, defined in section 3.4.1. All videos, using a consistent QoS (25 fps, 8-bit colour depth and a consistent objective sound quality), were embedded in an Internet Explorer window and shown via a projector system on to a large screen. The audio stream for all clips was consistently fed through amplified speakers at all four corners of the room.

Each video was shown three times. During the first showing, participants were asked to assess : i) how enjoyable they found the video content, as well as ii) how informative did they perceived the video as being. Assessment was made using two 7-point Likert scales, with scores of 0 through 6 rated as '*not*', '*hardly*', '*slightly*', '*fairly*', '*reasonably*', '*very*' and '*completely*' respectively. During the second showing participants were asked to judge the perceived importance of dynamic (D), audio (A), video (V) and textual (T) information, with scores of 0 to 2 representing respectively *low*, *medium* and *high*. The D, A, V, T criteria was used to ensure consistency with characteristic weightings previously defined by Ghinea and Thomas (see Table 2.3). A pause was given after the second viewing to ensure that participants had adequate time to complete this section of the form. During the third showing, participants were asked to check the previous assessments and make any necessary corrections. After all the videos had been shown, users were invited to provide additional feedback of any problems that were identified.

3.4.2.1. Enjoyment and Informative

Results show that certain videos can be generalised as inherently informative, yet cannot be determined as significantly enjoyable. An ANOVA (Analysis Of Variance), with Video Clip as the independent variable and D, A, V, T, Informative and Enjoyment Ratings as dependent variables was used. Results showed that the video clip significantly impacts the user's perception of whether a video is considered as being informative {F(1,11) = 14.747 p<0.001}(see Figure 3.2b); video clip does not significantly affect user enjoyment rating (see Figure 3.2a). User level of enjoyment varies considerable between users, i.e. enjoyment is participant dependent. User informative ratings, however, vary between videos, but not significantly between users, i.e. it is video dependent. Our results support the use of the original video material, as user perception concerning the informative nature of different video clips was found not to be participant dependent, yet instead varies between different clips.

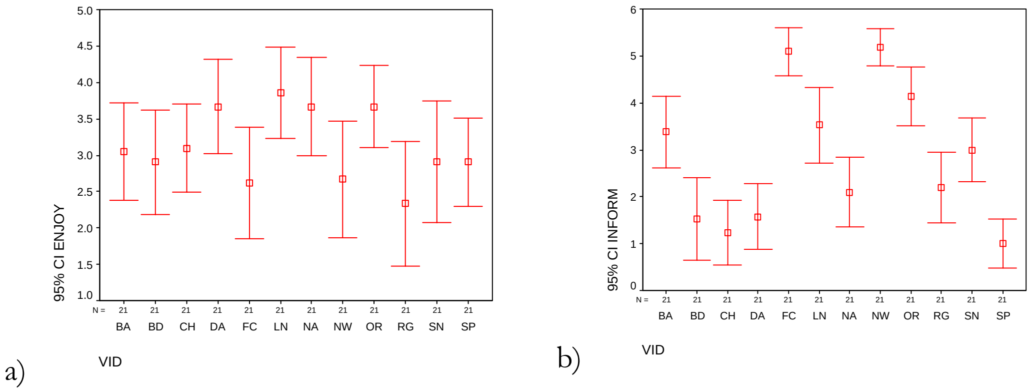


Figure 3.2: Mean and Standard Deviation for a) User Enjoyment b) User perception of Informative

The results of this study also highlight that the video being shown has a significant affect on the user perceived level of dynamic, audio and textual information. Although video clip was found not to significantly affect V rating, a number of participants identified that they failed to understand the difference between video and dynamic information. Importantly, results showed that variation occurs in the informative rating of both video and audio streams as a result of video type.

An ANOVA was used, with age as the independent variable (0-15, 16-30, 31-45, 46-60, 60+), and D, A, V, T, Informative and Enjoyment Ratings as dependent variables. Results showed that participant age significantly impacts both level of enjoyment {F(1,4) = 4.374 p=0.002} and perceived informative rating {F(1,4) = 3.367 p=0.011}. Results show that participant age significantly impacts the perceived level of audio (A) {F(1,4) = 0.383 p=0.002} and textual (T) {F(1,4) = 4.042 p=0.004} information. Post-Hoc Tukey-Tests, however, showed that this significance was only demonstrated in the over 60's group. We believe that this result is as a direct result of hearing loss. Interestingly, [GUL03] showed an identical result for users with reduced hearing level, who become increasing dependent on textual information to contextually understand the video content. When we statistically combined the impact of age with video type, no significant changes occurred in defined level of D, A, V and T. Consequently, as long as experimental participants are less than 60 years of age, consistent varied information definition is valid when using the original video material.

3.4.2.2. Comparing Characteristic Weightings

Ghinea and Thomas [GHI98] defined the original video material using characteristic weightings, with scores from 0 to 2 representing respectively low, medium and high level of importance. By averaging user feedback, we were able to determine user-defined importance ratings, as shown in Figure 3.3. By rounding these averages we were able to identify how closely our test matches the characteristic weightings, originally used by Ghinea and Thomas (see Table 2.3).

A Pearson Correlation was used to compare the average user feedback ratings and Ghinea and Thomas original characteristic weightings. Results showed a significant correlation value {r (48) =0.65, P<0.001}, which implies that user definition of quality is significantly similar when shown consistent video. In fact user definition was observed as being close to the rounding threshold for most cases where results differed (see Table 3.1).

The video material, originally used by Ghinea and Thomas [GHI98], possesses a video-dependent information level, which assuming experimental participants are less than 60 years of age, shows significant variation in level of dynamic, audio, video and textual data. Moreover, we have shown that characteristic weightings are consistently determined, across studies, which implies consistent user perception.

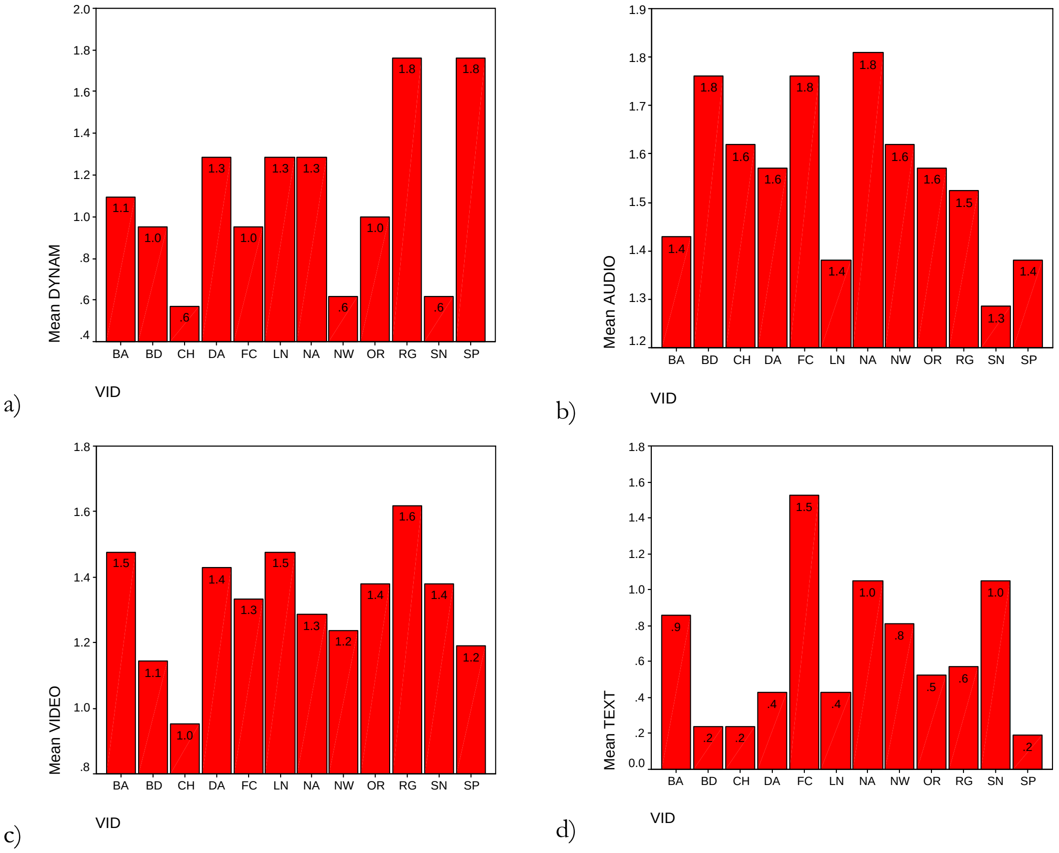


Figure 3.3: Average a) Dynamic- b) Audio- c) Video- d) Textual- Ratings.

Table 3.1: Difference between user rating and Ghinea and Thomas characteristic weightings.

	Dynamic (D)		Audio (A)		Video (V)		Textual (T)	
	Rating	Ghinea	Rating	Ghinea	Rating	Ghinea	Rating	Ghinea
BA	1	1	1.4	2	2	2	1	1
BD	1	1	2	2	1	1	0	0
CH	0.6	0	2	2	1	1	0	0
DA	1	1	1.6	1	1.4	2	0	0
FC	1	0	2	2	1.3	2	2	2
LN	1	1	1.4	2	2	2	0	0
NA	1	1	2	2	1.3	2	1	2
NW	0.6	0	2	2	1.2	2	1	1
OR	1	0	2	2	1.4	2	.5	0
RG	2	2	1.5	1	2	2	1	1
SN	0.6	0	1	1	1	1	1	2
SP	2	2	1	1	1.2	2	0	0

3.4.3. Experimental Questionnaire

Previously, we justified using video material initially defined by Ghinea and Thomas [GHI98]. To ensure consistency between QoP experiments, similar questionnaire material should be used. In our study the dynamic and video information sources, originally used by Ghinea and Thomas, were combined. Accordingly, slight differences were made to the experimental questionnaire. QoP-IA questions can be answered if and only if the user assimilates information from specific information sources. As the emphasis of information assimilation varies between different videos (see Table 3.1), the importance of gaining feedback from specific information sources also varies considerably across the clips. Accordingly, the number of QoP-IA questions, relating to different information sources, also varies (see Table 3.2). For full details of the experimental questionnaire, see Appendix A.

Table 3.2: QoP Information Distribution - Video, Audio and Textual.

VIDEO	**VIDEO (answers requiring V)**	**AUDIO (answers requiring A)**	**TEXT (answers requiring T)**
BA	7	2	0
BD	10	1	0
CH	11	0	0
DA	8	4	1
FC	8	3	5
LN	10	1	0
NA	6	2	4
NW	1	11	3
OR	9	6	0
RG	9	3	3
SN	8	1	4
SP	9	1	0

To analyse results statistically, we used SPSS (Statistical Package for the Social Sciences - Version 11.5), which is a data management and analysis product produced by SPSS, Inc. in Chicago, Illinois. Among its features are modules for statistical data analysis, including descriptive statistics such as plots, frequencies, charts, and lists, as well as sophisticated inferential and multivariate statistical procedures like analysis of variance, factor analysis, cluster analysis, and categorical data analysis. As well as being user accessible, SPSS is particularly well suited to survey-based (questionnaire-based) research, which is why it has been used for our analysis. A number of statistical tests are used in our studies. These include: ANOVA (ANalysis Of Variance) and MANOVA (Multiple ANalysis Of Variance) tests (both supported by Post-Hoc Tukey tests), as well as Willis K-independent Non-parametric tests (depending on the ANOVA homogeneity of variables), which were used to determine the impact of parameter variance on user QoP. Pearsons and Kendall's tau-b Bivariate Correlations, were used to identify correlations between data, i.e. comparison of eye tracking data, as a result of QoS variation. For all tests a result was considered to be significant if p<=0.05. This implies that the mean of a specific data set is greater / less than two standard deviations from the overall mean, as a result of a specific variable variation - approximately 5 percent of all samples. In addition to statistical analysis, creation of data images and comparison of image content has been used to analyse frame content and facilitate eye-tracking data evaluation.

3.5. Eye Tracking

Eye-tracking will be employed in our work to help identify how gaze disparity in eye-location is affected at the network-level. By continuously monitoring user focus we aim to gain a better understanding why people do not notice obvious cues in the experimental video material. Eye-tracking will be measured at the network-level, however, due to the complexity of eye-tracking data, will be analysed separately to QoP data. Eye-tracking data will be used at the media-level to manipulate video content, yet no monitoring of user eye-gaze location will be made at the media-level. Although an eye-tracker will be used as a display device, no monitoring of user eye-gaze location data will be made at the content-level.

3.5.1. Understanding Eye Movement

If visual acuity were evenly distributed across the surface of the retina, human spatial acuity would be poor. Accordingly, cones are concentrated in the central area of the retina, in a circular area called the fovea, extending over just 2° of the visual field. Consequently, to ensure high acuity, the eye must be moved in such a way that the target object can be inspected with a higher acuity, by foveating the object, i.e. moving the eye to ensure that the *Region of Interest* (RoI) is projected on the centre of the fovea. There are seven specific patterns of eye movement considered by eye-tracking devices [ENG95a]:

- **Saccades:** *Saccades* (or saccadic movement) is the principal method for moving the eyes within the visual field. Saccades are sudden, ballistic movements of the eyes, taking between 130 and 420ms. Saccades can be initiated voluntarily, however, once initiated they cannot be changed [ENG95b].

- **Pursuit:** Smooth pursuit is an even shifting of the eye to keep a moving object foveated, i.e. at the centre of the fovea. It cannot be induced voluntarily, but requires a moving object in the visual field [ENG95b].

- **Convergence:** Humans have stereoscopic (3D) vision. This means that we have the ability to determine distance as our two eyes are separated by a few centimetres (termed as the disparity)**.** There is an inverse relationship between disparity and the depth in the scene; disparity will be relatively large for points in the scene that are near to us and relatively small for points that are far away. To focus both eyes at one specific object, convergence occurs to shift the object in the visual field. Convergence is therefore the motion of both eyes*,* relative to each other, to ensure that an object is foveated by both eyes when the distance between the observer and the object changes. This movement can be voluntarily controlled, but is normally the result of a moving stimulus [ENG95b].

- **Rolling:** Rolling is a rotational motion around an axis passing through the fovea and pupil. Rolling is involuntary, and is especially influenced by the angle of the head [ENG95b].

- **Nystagmus:** Nystagmus occurs as a response to viewing a repetitive moving object. It consists of a '*smooth pursuit*' motion in one direction, followed by a fast motion in the opposite direction to select a new Region of Interest (RoI) [ENG95b].

- **Drift and microsaccades:** Drift and microsaccades are involuntary movements that occur when the eye is fixated, i.e. is focusing on a single object, and consists of a slow drifts followed by corrective micro-saccades [ENG95b].

- **Physiological nystagmus:** Physiological nystagmus is high-frequency oscillation of the eye, which moves the image on the retina, thus reducing retinal masking. If any image is artificially fixed on the retina, visual masking occurs and the image is perceived to disappear. Physiological nystagmus causes the retinal image to move approximately the distance between two adjacent foveal cones every 0.1 seconds. Physiological nystagmus occurs during a fixation period, is involuntary and moves the eye approximately 1° [ENG95b].

3.5.2. Defining the Ideal Eye-Tracker System

Scott and Findlay [SCO92] stated a number of usability requirements that define the ideal eye-tracking device. Accordingly, the ideal eye-tracking device should [ENG95c]:

- Offer an unobstructed field of view with good access to the face and head.
- Make no contact with the subject.
- Meet the practical challenge of being capable of artificially stabilising the retinal image if necessary.
- Possess an accuracy of at least one percent or an arc minute (a unit of angular distance equal to a 60th of a degree), i.e.: an eye-tracker would not give a 10° reading when it is truly 9°.
- Offer a resolution of 1 arc minute, and thus be capable of detecting the smallest changes in eye position.
- Offer a wide dynamic range from one arc minute to 45° for eye position and be capable of measuring changes in eye velocity from 1 to 800 arc minutes per second.
- Possess a real-time response, to allow physiological interaction.
- Measure all three degrees of angular rotation of the eye.
- Facilitate binocular data recording.
- Be compatible with head and body recordings
- Finally, be capable of being used on a variety and range of participants.

3.5.3. Eye Tracking Techniques

Although all requirements are ideally desirable, it is accepted that not all are practical prerequisites for an acceptable eye tracking system. There are currently several approaches to sensing eye movements [YOU75], including:

- Measuring the reflection of light that is shone onto the eye.
- Measuring the electric potential of the skin around the eyes.
- Applying a special contact lens that facilitates tracking of the eye position.

Moreover, eye-tracking equipment measures either the saccadic movement of the eye or the point of fixation and are accordingly categorised as being either fixation or saccadic pickers [CLA99].

3.5.3.1. Techniques Based on Reflected Light

There are five tracking techniques that facilitate light reflected by the eye: limbus tracking, pupil tracking, corneal and pupil reflection relationship, corneal reflection / eye image using an artificial neural network, and finally Purkinje image tracking [ENG95d; HUT89; POM93; RAD94].

Limbus Tracking: The limbus is the boundary between the white of the eye (sclera) and the coloured iris of the eye (see Figure 2.2). Due to this light / dark division, this boundary can easily be detected and tracked. This technique is based on the position and shape of the limbus relative to the head, so either the head position must be maintained or the apparatus must be fixed to the user's head [SCO93].

Pupil Tracking: Tracking the direction of pupil is similar to limbus tracking, however, the boundary between the pupil and the iris is used. Once again, the apparatus must be held completely still in relation to the head. Advantages of pupil tracking over limbus tracking include:

- the fact that the pupil is far less covered by the eyelids, limiting error during vertical and horizontal tracking.
- the fact that the border of the pupil is sharper than that of the limbus, which yields a higher resolution.

The disadvantage of pupil tracking is that the difference in contrast is lower between the pupil and the iris than between the iris and the sclera.

Corneal Reflection and Pupil Reflection Relationship: When (infrared) light is shone into the user's eye, several reflections occur on the boundaries of the lens and cornea (see Figure 2.2). By analysing the relative position of these reflections the direction of gaze can be calculated.

The problems associated with this technique are primarily related to getting a good view of the eye. Head movement can put the video image out of focus, or even move the eye out of view of the camera [JAC01; SCO93].

Corneal Reflection and Eye Image using an Artificial Neural Network: Using a digitised image of the user's head (a wider angle than in other techniques) the system finds the right eye of the user, by using a light to create a glint on the eye (a small very bright point surrounded by a darker region). The system then extracts a (40 x 15 pixel) rectangular video image, centred on the glint [BAL94].

An artificial neural network is subsequently used to identify the direction of eye-gaze. The main advantage of this technique is that the wide angle of the image allows user head mobility. Disadvantages include:

- The artificial neural network requires substantial training for each user.
- The general accuracy of such systems is not currently as good as other techniques.

Purkinje Image Tracking: Using reflection of light, separate reflective images can be formed called the 'Purkinje images'. Corneal and pupil reflections are known as second and third Purkinje images. Similar to the corneal and pupil reflection relationship, relative positions of first and fourth Purkinje images can be used to calculate position of gaze [MUL93]. Although accurate, the fourth Purkinje image is weak, so the surrounding lighting must be heavily controlled if accuracy is to be maintained [CLE92].

3.5.3.2. Techniques Based on Electric Skin Potential

Electro-OculoGraphic or ElectroNystagmoGraphic potential (EOG / ENG) tracking is based on the existence of electrostatic fields around the eye, which are created by muscles during the movement of the eye. By recording small differences in the skin potential around the eye, the position of the eye can be detected. This is done using electrodes placed on the skin around the eye and enables EOG tracking techniques, which measure approximately ±70° of the visual field [ENG95e; GIP93; GIP96; SCO93].

3.5.3.3. Techniques Based on Contact Lenses

By making the user wear a special contact lens, it is possible to make accurate recordings of the direction of gaze. Eye tracking techniques that facilitate contact lenses are extremely intrusive and are restricted to laboratory conditions. Two specific eye-tracking methods use contact lenses [ENG95f]:

- by engraving one or more plane mirror surfaces on the lens, and calculating the position of the eye from reflected light beams.
- by implanting a tiny induction coil into the contact lens, and using a high-frequency electro-magnetic field placed around the user's head, to determine the exact position of the lens.

3.5.4. Relating Eye-tracking and QoS

The eye naturally focuses on areas that are most likely to be informative [MAC70]. Mackworth and Bruner [MAC70] studied the eye movement of participants while looking at blurred pictures. The visual area was divided into 64 squares, each with an informative weighting. The most informative areas attracted more fixations [MAC67; MAC70]. Mackworth and Morandi also noted that informative areas are identified within the first two seconds of observation [MAC67], a conclusion that has been reported in other studies of eye movement [DEG67; YAR67]. Consequently, since monitoring eye movements offers considerable insight into visual perception, as well as the associated attention mechanisms and cognitive processes, it is a valid way of determining factors that affect user perceptual processes.

Consequently, eye tracking is being employed in the design of user interfaces, as an efficient interface ensures, for instance, that commonly-used controls are located in areas where the eyes' gaze is most likely to rest [PAR00], and that eye movement between these controls is minimal. Additionally, eye-based interfaces also help users (especially disabled) to execute interface input actions, such as menu selection [BYR99], eye-typing [SAL99], and even mouse clicking, through the development of an 'eye-mouse' [LAN00]. Web design guidelines based on results obtained using eye tracking technology have also been elaborated and are being used by commercial web designers to write more effective web pages [NIE00]. Eye tracking is also currently being used in virtual reality-based education and training, ranging from such diverse topics as aircraft inspection [DUC00] to driving [SCI00]. However, in the context of this paper, we are interested in the relationship between eye movement and network-level technical variation.

3.5.5. Implementing Eye-tracking

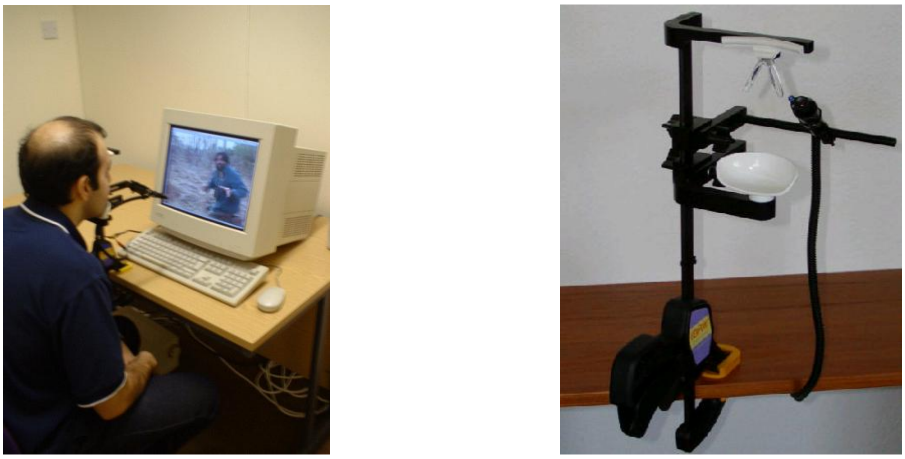
Variations in different eye-tracking systems help facilitate a wide range of functionality. Eye-tracking systems can be used as a data-gathering device or can provide the user with interactive functionality [ISO00; REI02]. Depending on the equipment, eye-tracking devices can be considered as either intrusive or non-intrusive in nature [GOL02] and can be developed as either pervasive [SOD02] or standalone systems. Level of immersion, perceived whilst using eye-tracking equipment, may be high [HAY02] or low [PAS00], depending on the specific equipment type. Accordingly, proper consideration must be given to the eye-tracking device used in our investigation, to ensure that effective experimental method and data collection is achieved.

The two important practical issues influenced the choice of eye-tracking system, used in this research. These issues were: the budget and the required system functionality.

1) **Cost:** £5000 was allocated in order that we may purchase an eye-tracking system, for use in our study.

2) **Functionality:** The majority of eye tracking research relies on static visual stimuli, e.g. a picture or a static web page. Consequently, only a limited number of systems facilitate the use of video stimuli and appropriate data storage. Although eye-tracking systems have been developed for real-time manipulation of video, i.e. dedicated gaze-contingent display systems [PAR00], these systems were well outside our budget.

We chose to purchase an Arrington Research ViewPoint EyeTracker (a Macintosh based system that uses an infrared camera to provide corneal / pupil reflection eye tracking – see Figure 3.4a) in combination with QuickClamp Hardware (see Figure 3.4b). See Table 3.3 for a detailed technical specification of ViewPoint Eye-Tracker.



a) b)

Figure 3.4: a) Power Mac G3 (9.2) ViewPoint EyeTracker, used in combination with QuickClamp Hardware b).

Table 3.3: Technical Specification of ViewPoint Eye-Tracker

Accuracy	Approximately 0.5° - 1.0° visual arc
Temporal resolution	Maximum resolution - 30 Hz
Visual range	Horizontal:+/-44°of visual arc Vertical: +/- 20 ° of visual arc
Calibration	Calibration is required only once per subject. New subject set-up time between 1-5 minutes. Calibration settings can be stored and reused each time a subject returns.
Blink suppression	Software contains automatic blink detection and suppression.
Data recorded	(see Appendix C)

The Arrington Research ViewPoint EyeTracker facilitated streaming video in the eye tracking stimulus windows and was priced within our budget. Eye-tracking data output includes: X coordinate values, Y coordinate values and timing data (a delta time that represents the time {ms} between samples) - see Appendix C. X and Y coordinate values (ranging 0-10000) were defined automatically by the ViewPoint EyeTracker system, and represented the minimum and respectively the maximum horizontal and vertical angular extent of eye movements on the screen, from the top left corner (0,0) to the bottom right corner (10000, 10000). In order to simplify data synchronisation between participants, eye-tracking data was sampled at 25Hz for all participants, corresponding to the maximum frame rate of the experiment material. The output format of the ViewPoint EyeTracker data was far from ideal. Consequently, substantial difficulty was faced: i) combining data from multiple videos in a single participant data file; ii) cleaning / removing unwanted data from participant data files (see Appendix C); and iii) synchronising eye tracking data from multiple participants. We solved these problems by developing state logic scripts (see Appendix D), which: i) conjoined multiple videos into a single presentation, thus facilitating a single data file for each participant; and ii) allowed us to flag essential data and timing points in the output data file, which aided subsequent analysis, cleaning and synchronisation of data.

Dedicated software was written (see Appendix E), which automatically analysed eye tracking output data files, and using implemented state logic flags, extracted, ordered, and synchronised appropriate eye-tracking data for each participant in turn. This software used a three-stage approach: firstly, essential data was imported from the participant's eye tracking data file; secondly, data and timing flags were interpreted by the system, allowing the eye-tracking data to be synchronised and appropriately stored in a dedicated object-oriented data structure (see Figure 3.5); and finally, as required, the contents of the data structure were exported (see Appendix C), which allowed us to re-load consistent data when required.

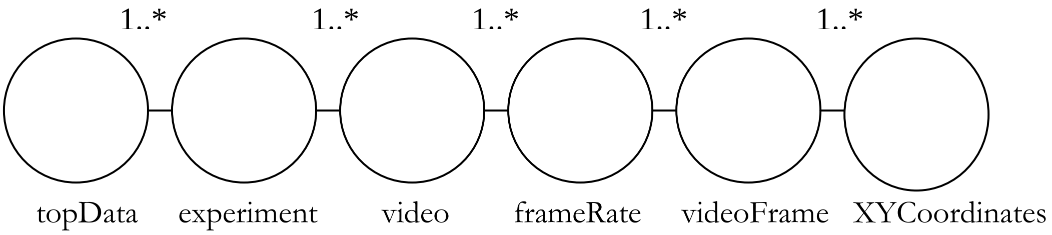


Figure 3.5: Object-Orientated Data Structure, used to store participant eye-tracking data (See Appendix E).

Consequently, although the data output of the ViewPoint EyeTracker system was far from ideal, additional flagging of essential events in the data file facilitated the effective cleaning and synchronisation of eye-tracking data from multiple video and participants.

3.6. Applied Experimental Method

3.6.1. Experimental Participants

Participant numbers were determined by two factors: the number of variable factors in each experiment and the practical availability of subjects. Each participant that was used in our experiments had never participated in a QoP experiment before, thus minimising the existence of participant pre-knowledge. Participants used in our experiments were taken from a range of different nationalities and backgrounds – students, clerical and academic staff, white collar workers, as well as a number of retired persons. All participants, however, spoke English as their first language, or to a degree-level qualification, and were computer literate.

In previous studies (see Table 2.4), Ghinea and Thomas [GHI98] used 30 participants to measure the impact of both frame-rate and video content on user perception. Procter et al. [PRO99] used 24 participants to measure the impact of both network load and video content on user perception. In our study we matched the participant numbers used in previous perceptual studies.

Importantly, the majority of statistical analysis in our work relies on one-way ANOVA F tests. Assuming the sample is not unbalanced, one-way ANOVA F tests will not be seriously affected unless the sample sizes is less than 5, or the departure from normality is extreme. Accordingly, despite within-measures variation, we aimed to have a minimum sample size of 6. Although this was achieved for the majority of our studies, practical limitation on the number of available participants meant that a minimum sample size of four was achieved when analysing the impact of frame rate in objective 3 (see chapter 6). Using a minimum sample size of four reduces the reliability of results concerning the impact of frame-rate on user perception of quality, it does not impact results relating to device variation or video clip type.

3.6.2. Experimental Process

All experiments used in our work followed a similar consistent experimental process. To avoid audio and visual distraction, a dedicated, uncluttered room was used throughout all experiments. All participants were asked a number of short questions concerning their sight, which was followed by a basic eye-test to ensure that all participants were able to view menu text on the screen. This was specifically important for those using the eye-tracking device, as participants were not able to wear corrective spectacles for the duration of the experiment. Participants were informed that after each video clip they would be required to stop and answer a number of questions that related to the video clip that had just been presented to them. To ensure that participants did not feel that their intelligence was being tested it was clearly explained that they should not be concerned if they were unable to answer any of the QoP-IA questions.

After introducing the participant to the experiment, the appropriate experimental software and video order were configured. In the case of the participants using the eye-tracker, time was taken to adjust the chin-rest, infrared red capture camera and software settings to ensure that pupil fix was maintained throughout the user's entire visual field. When appropriate calibration was complete, the participant was asked to get into a comfortable position and in the case of the eye-tracker place his/her chin on the chin-rest. The correct video order was loaded (see experimental chapters for details) and the first video was displayed.

The content of the videos used in our experimental presentations was manipulated to simulate specific quality parameter variation (see experimental chapters 4, 5 and 6 for details concerning video manipulation). Due to the reduced bandwidth requirement and increased perceptual impact of corrupted audio, the audio stream will not be manipulated in our research. By purely manipulating video content we minimise the number of variables that impact the user's perception of quality.

After showing each video clip, the video window was closed and the participant was asked a number of QoP questions relating to the video that they had just been shown. QoP questions (as described in Appendix A) were used to encompass both QoP-IA and QoP-S (QoP-LOE and QoP-LOQ) aspects of the information being presented to the user. The participant was asked all questions aurally and the answers to all questions were noted at the time of asking.

Once a user had answered all questions relating to a specific video clip, and all responses had been noted, participants were presented with the next video clip. This was done for all 12 videos, independent of the display device.

3.7. Conclusion

In summary, this chapter considered and justified the use of structured laboratory experiments. It also described and justifies the use of adapted QoP (Quality of Perception), the experimental material, the experimental questionnaire and analysis method that shall be used, in order to achieve the defined research aim and objectives. Moreover, we provided the reader with a background concerning eye-tracking, justify why eye tracking will be considered in our experiment, and how it has been implemented in our work. In chapters 4, 5 and 6 we measure the user-perspective concerning network-, media-, and content-level quality parameter variation, as defined by our research objectives.

CHAPTER 4

Network-Level Quality Parameter Variation

4.1. Introduction

In our study, we aim to extensively consider the user-perspective regarding multimedia quality, by adapting relevant technical- and user-perspective parameters at all quality abstractions: network-level (technical-perspective), media-level (technical- and user-perspectives), and content-level (technical- and user-perspectives). Consequently, in this chapter, we intend to consider objective 1, which concerns the impact of network-level parameter variation (delay and jitter) on the user's perception of multimedia quality. In addition, eye-tracking will be employed in our work to help identify how gaze disparity in eye-location is affected at the network-level. As stated in chapter 3, by continuously monitoring user focus we aim to gain a better understanding why people do not notice obvious cues in the experimental video material. Eye-tracking will be measured at the network-level, however, due to the complexity of eye-tracking data, will be analysed separately to QoP data.

Although the impact of delay and jitter has been considered by other authors, these studies fail to measure the user-perspective considering both level of user understanding (information assimilation) and user satisfaction (concerning both of the video QoS {media-level}and enjoyment regarding the content of the video {content-level}).

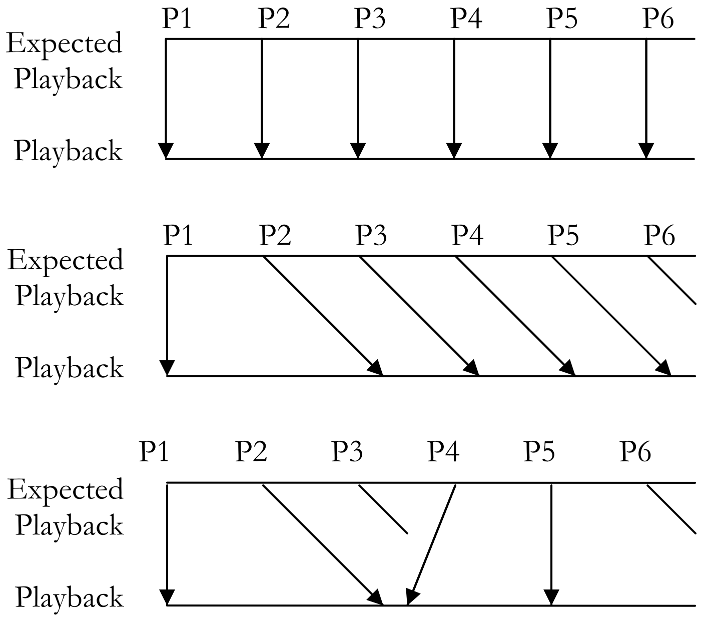
The structure of this chapter is as follows: we start this chapter, in section 4.2, by specifically describing the jitter and delay concepts and provide a summary of the relevant studies relating to their impact on perceptual quality. In section 4.3, we discuss how video material was manipulated to simulate delay and jitter, as well as information concerning the experimental process used specifically to measure the impact of network-level quality parameter variation on user perception of multimedia quality. In section 4.4, we consider our results, and lastly, in section 4.5, conclusions are discussed.

4.2. Delay and Jitter in Networked Multimedia

There are three measures that determine network-level quality in most digital distributed video applications: delay, packet loss and jitter [CLA98]. In this study we aim to manipulate delay and jitter. Loss was not considered in this study, as the impact of loss was investigated by Ghinea and Thomas [GHP00] in light of original QoP definition. Accordingly, it seems a misuse of resource to duplicate this work.

Delay: Delay is the time taken by a packet to travel from the sender to the recipient. A delay is always incurred when sending distributed video packets, however the delay of consecutive packets is rarely constant [WAN01] (See Figure 4.1); a phenomenon that gives rise to jitter.

Jitter: Jitter is the variation in the period of delay [WAN01] (See Figure 4.1).



No Delay or Jitter

Delay: Delay in P2 causes delay in subsequent playback of other packets.

Jitter: Positive jitter (P2) and negative jitter (P4) cause loss of P3

Figure 4.1: The effect of Delay and Jitter on video playback.

In Chapter two, we make reference to a numerous studies concerning the perceptual impact of delay and jitter. In summary, these studies showed that:

- Jitter degrades video quality nearly as much as packet loss [CLA99].
- The presence of even low amounts of jitter or packet loss results in a severe degradation in perceptual quality. However, higher amounts of jitter and packet loss do not degrade perceptual quality proportionally [WIJ96].

- Perceived quality of low temporal aspect video is not impacted in the presence of jitter as much as video high temporal aspect [CLA99; KAW95; STE96; WIJ96].

- There is a strong correlation between the average number of quality degradation events (points on the screen where quality is affected) and the average user quality rating recorded. This suggests that the number of degradation events is a good indicator of whether a user will like a video presentation effected by jitter and packet loss [CLA99; WIJ96].

- Momentary rate variations in the audio stream, although initially considered as being amusing, were soon deemed to be annoying in experimental studies. This resulted in participants concentrating more on the audio defect, rather than the audio content [CLA99].

The presence of video delay especially impacts the synchronisation of audio and video streams. Steinmetz [STE96] summarises the minimal synchronisation errors that have been found as perceptually acceptable (see Table 4.1).

Table 4.1: Minimal noticeable synchronisation error [STE96].

Media		Mode, Application	QoS
Video	Animation	Correlated	±120ms
	Audio	Lip synchronisation	±80ms
	Image	Overlay	±240ms
		Non-overlay	±500ms
	Text	Overlay	±240ms
		Non-overlay	±500ms
Audio	Animation	Event correlation (e.g. dancing)	±80ms
	Audio	Tightly coupled (stereo)	±11µs
		Loosely coupled (e.g. dialogue with various participants)	±120ms
		Loosely coupled (e.g. background music)	±500ms
	Image	Tightly coupled (e.g. music with notes)	±5ms
		Loosely coupled (e.g. slide show)	±500ms
	Text	Text annotation	±240ms
	Pointer	Audio relates to showed item	(-500ms, +750ms)

4.3. Experimental Approach

In this section, we consider issues relating to the experimental approach, which was used in our study to measure the impact of network-level quality parameter variation on user perception of multimedia quality.

4.3.1. Creating Jitter and Delay Video Material

In this section, we consider how the original MPEG videos (see Appendix B) were manipulated to simulate network-level delay and jitter quality parameter variation. MPEG (Moving Picture Experts Group) video compression is the internationally recognised standard for motion picture compression and is used in most current and emerging digital technologies. MPEG facilitates the fact that consecutive frames are similar in nature by "losing" repetitive information during compression and encoding. This method, known as 'prediction with movement compensation', allows MPEG decompression to deduce the content of some pictures from references to preceding and subsequent frames, thus minimising the transfer of data. Three frame types exist (I, P and B), with an order as demonstrated in Figure 4.2.

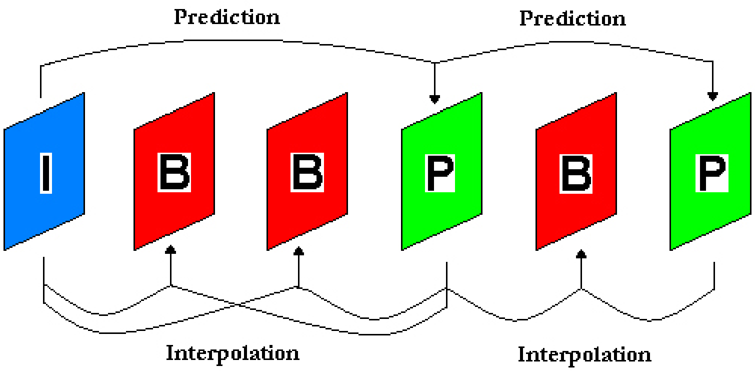


Figure 4.2: Concatenation of I, B and P picture in MPEG video.

I (intra) frames are compressed, yet have no reference or dependence on the contents of other frames. An I frame contains all of the information required to reconstruct an image. Consequently I frames are used as an 'entry point' for video playback. P (predicted) frames are compressed, with reference to information in preceding I or P frames. Unfortunately, because P frames are predicted, leading to error in the motion compensation, it is not possible to extend the number of P frames between two I frames. Accordingly, B (bi-directional or bi-directionally predicted) frames are placed between I and P pictures and uses the information from I and P frames to interpolate the difference. There are currently five distinctive MPEG formats [CHI04]:

- MPEG-1: Coding of moving pictures and associated audio for digital storage media at up to about 1.5 MBit/s.
- MPEG-2: Generic coding of moving pictures and associated audio.
- MPEG-4: Very low bitrate audio-visual coding.
- MPEG-7: Multimedia content description.
- MPEG-21: Open framework for multimedia delivery and consumption.

In order to ensure consistency with the original MPEG-1 videos (352*288 pixels) [GHI98], our work is interesting in MPEG-1. Accordingly, to simulate delay and jitter we artificially manipulated skew between audio and video media streams. We manipulated video so that the number of delay and jitter errors equalled 2% the number of video frames, which corresponds to one video error every two seconds (the minimum time taken to identify Regions of Interest (RoI) in a visual stimuli [MAC67; DEG67; YAR67]). Consequently, to simulate accumulated video delay, after every 50 video frames a single video frame was repeated, i.e. for 50 original frames, 51 were shown. At no point was the audio manipulated. As a consequence of duplicated video frames, the manipulated delay video was 2% longer than the audio stream. To simulate video jitter, which is the variation in delay, a number of jitter points were simulated that was equal to 2% the number of video frames, e.g. for a 918 frame video, 18 separate jitter points were simulated. The location of jitter points was randomly defined. The direction (+/-) and amplitude of each video skew (0 - 4 frames) was also randomly defined, however, minute adjustments were made to ensure that the net delay was equal to zero, i.e. the first and last video frame synchronised with the audio stream. Randomly sized video skew (0 - 4 frames) was used to ensure variation in jitter, ranging from 0ms to 160ms, which represents a maximum skew equal to two times the minimal noticeable synchronisation error between video and audio media - see Table 4.1. Manipulation of delay and jitter videos was achieved by manually shifting video skews in Adobe Premier 5. Finally, manipulated video were encoded as (352*288 pixels) MPEG-1 videos to ensure QoS consistency with the original videos. Videos were cast as 5, 15 and 25 fps, which allowed user perception to be measured as a result of both quality variation and frame-rate variation. Video variation therefore includes: tuple{fps, control}(5, 15, 25); tuple{fps, delay} (5, 15, 25); and tuple{fps, jitter} (5, 15, 25). Accordingly, nine combinations of video variation existed (*Video Quality Types*), as shown in Table 4.2.

Table 4.2: Videos used in Control / Delay / Jitter network-level perceptual experiments.

Video Quality Types	Description
O5	No Delay or Jitter (5fps)
O15	No Delay or Jitter (15fps)
O25	No Delay or Jitter (25fps)
J5	Jitter (5fps)
J15	Jitter (15fps)
J25	Jitter (25fps)
D5	Delay (5fps)
D15	Delay (15fps)
D25	Delay (25fps)

4.3.2. Experimental Variables

Three experimental variables were manipulated in this chapter: simulated network-level quality variation, multimedia video frame rate and multimedia content. Accordingly, original, delay and jitter video conditions were considered in our experiments, and three multimedia video frame rates: 5, 15 and 25 fps. As far as multimedia content is concerned, 12 video clips (see Appendix B) were considered in our experiments.

4.3.3. Experimental Methodology

4.3.3.1. Participant Distribution

Participants were aged between 18 and 57. To measure the impact of network-level quality parameter variation (jitter and delay) on user perception, 108 participants were evenly divided into three experimental groups (C, J and D), which related to the perceptual impact of control, jitter and delay videos respectively. Participants in each group (36 participants in each of the three groups) were divided into three sub-groups (1, 2 and 3), each containing 12 participants, which were used to distinguish the impact of viewing order and frame rate – see Figure 4.3.

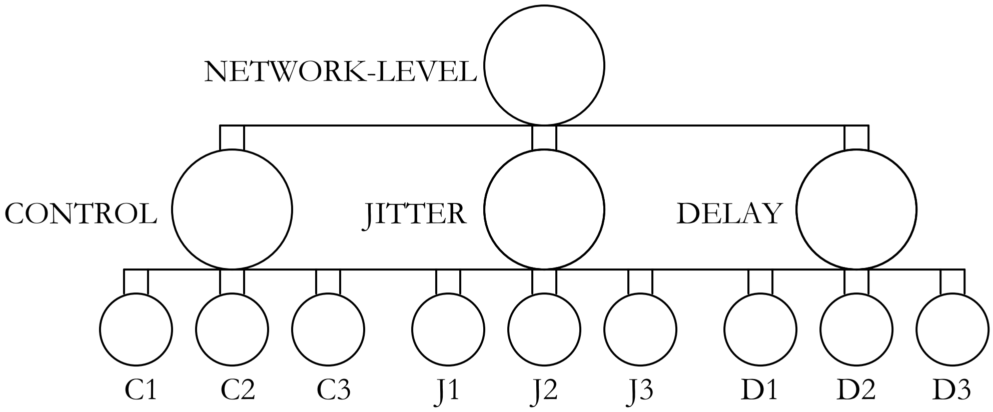


Figure 4.3: Participant distribution in order to measure impact of network quality parameter variation (Delay and Jitter).

In each experimental sub-group (e.g. C1, C2, C3, J1, etc.), a within-subjects design was used, where participants viewed each of the 12 video clips in turn at one of three pre-recorded frame rates (5, 15 or 25fps). Thus, each participant viewed four video clips at 5 fps, four at 15 fps, and four at 25 fps. In order to counteract order effects, the video clips were shown in a number of order and frame-rate combinations, defined by the experimental sub-group name, e.g. J3 sub-group participants (see Figure 4.3) viewed videos with frame-rates as defined by column 'Content 3' (see Table 4.3).

Table 4.3: Frame-rate order for Control, Jitter and Delay sub-groups.

Video	Order	Content 1	Content 2	Content 3
BA	A	5	15	25
BD	A	25	5	15
CH	A	15	5	25
DA	A	25	15	5
FC	A	5	25	15
LN	A	5	15	25
NA	B	15	25	5
NW	B	5	25	15
OR	B	15	25	5
RG	B	25	5	15
SN	B	15	5	25
SP	B	25	15	5

Moreover, six participants from each sub-group viewed A-order videos first (Commercial through Space) and six watched B-order videos first (Pop-Space, Commercial-Documentary). To identify the order with which a participant viewed videos we can define a participant as being either A-order first (A) or B-order first

(B), i.e. a participant in group J3 were either J3a or J3b, depending on his/her viewing order. The large number of participants, and consequently participant viewing orders, was primarily used to ensure reliability of collected eye-tracking data.

4.3.3.2 Experimental Setup

To guarantee that experimental conditions remained constant for all control participants, consistent environmental conditions were used. An Arrington Research, Power Mac G3 infrared camera-based corneal / pupil tracking, ViewPoint EyeTracker was used, to extract eye-tracking data, in combination with QuickClamp Hardware (see chapter 3 for detailed technical information). The QuickClamp system is designed to limit head movement and includes chin, nose and forehead rests, whilst supporting the infrared camera. The position of nose and forehead rests remained constant throughout all experiments (45cm from the screen). The position of the chin rest and camera were, however, changed depending on the specific facial features of the participant. To limit physical constraints, except from those imposed by the QuickClamp hardware, tabletop multimedia speakers were used instead of headphone speakers. A consistent audio level (70dB) was used for all participants.

4.3.3.3 Experimental Process

An experimental process was used that is consistent with that defined in section 3.6. Participants wearing contact lenses were not asked to remove lenses, however, due to the eye-tracking device, special note was made and extra time was given when mapping the surface of the participant's eye to ensure that a pupil fix was maintained in the 'Eye Camera Window' throughout the entire visual field (see Figure 4.4). Once system configuration was complete, automatic calibration was made using a full screen stimulus window. However, point re-calibration was also used if an unexpected error, due to participant movement, e.g. a sneeze, caused a non-smooth pupil mapping in the eye-space window (see Figure 4.4).

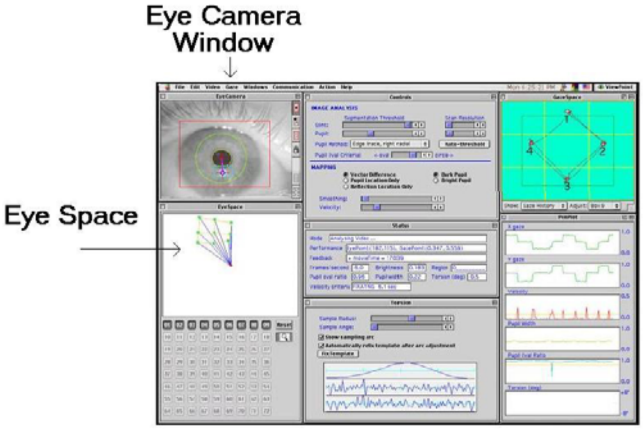


Figure 4.4: Layout of ViewPoint software - developed by Arrington Research Inc.

Once calibration of the eye-tracking system was complete, the appropriate presentation script (Control: C1A/B, C2A/B, C3A/B; Jitter: J1A/B, J2A/B, J3A/B; Delay D1A/B, D2A/B and D3A/B – see Appendix D) was loaded and the state logic script was incremented, which started the first video clip.

4.4. Results

4.4.1. Impact of Jitter and Delay on QoP-IA

QoP-IA is expressed as a percentage measure and reflects the level of information assimilated by the user from visualised multimedia content. An ANOVA (ANalysis Of VAriance) test, with video variation type (i.e. control, jitter and delay) as the independent variable and QoP-IA as a dependent variable, showed that video variation type has no significant impact on user QoP-IA (see Figure 4.5), which shows that the presence of delay and jitter do not impact a users ability to assimilate information.

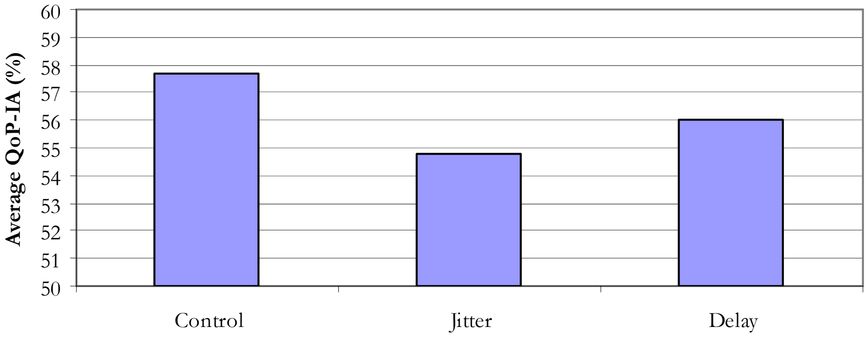


Figure 4.5: Impact of simulated network-level quality parameter variation on user QoP-IA.

Moreover, an ANOVA with video quality type (see Table 4.2) as the independent variable and QoP-IA as a dependent variable showed that quality type (see Table 4.3) does not impact user QoP-IA {F(1,8) = 1.311 p<0.234} (see Figure 4.6a). This implies that combined network-level quality parameter (jitter and delay) and frame rate variation does not significantly impact user QoP-IA.

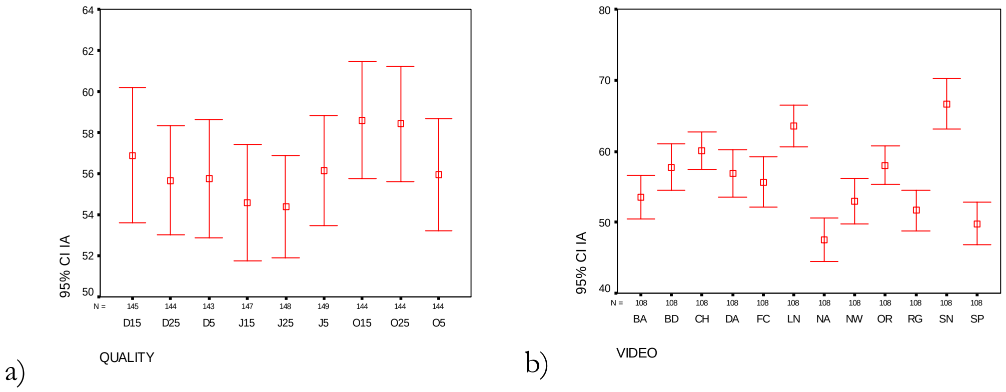


Figure 4.6: Impact of a) Quality type (see Table 4.3) and
b) Video type on user QoP-IA {Mean and St. Dev.}.

Interestingly, an ANOVA with video type (see Appendix B) as the independent variable and QoP-IA as a dependent variable, showed that video type significantly impacts user QoP-IA {F(1,11) = 12.700 p<0.001} (see Fig 26b). This finding supports the conclusion made in chapter 3, that the original videos used by Ghinea and Thomas conveyed a wide range of informational content.

4.4.2. Impact of Jitter and Delay on QoP-LoQ

An ANOVA with video variation type as the independent variable and QoP-LoQ as a dependent variable, showed that QoP-LoQ is significantly impacted by the presence of delay and jitter video variation {F(1,2) = 8.547 p<0.001} (see Figure 4.7). Moreover, post-Hoc Tukey-Tests showed a significant difference between the perceived QoP-LoQ for control and jitter {p=0.001}, as well as control and delay videos {p=0.002}. Results show that the presence of either jitter or delay causes a drop in user QoP-LoQ, which justifies the use of QoP-LoQ in context of this study. Moreover, results show that participants can effectively distinguish between a video presentation with and without error. This finding supports [WIJ96], who showed that the presence of even low amounts of network-level error results in a severe degradation in perceptual quality. It is therefore essential to identify the purpose of the multimedia presentation when defining appropriate network QoS provision, e.g. applications relying on user perception of multimedia quality should be given priority over and above purely educational applications.

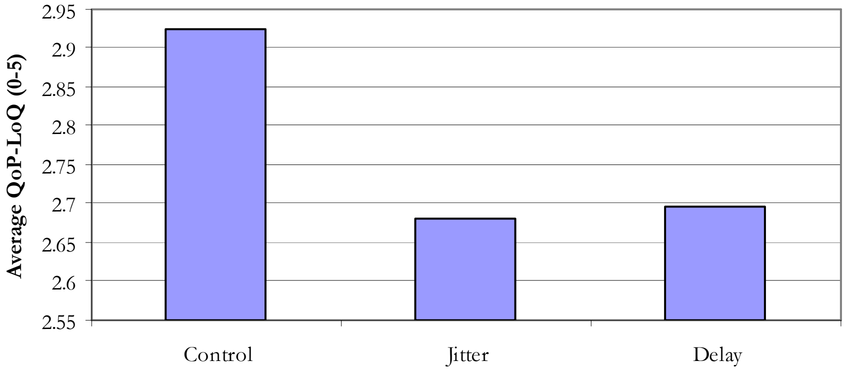


Figure 4.7: Impact of simulated network-level quality parameter variation on user QoP-LoQ.

An ANOVA (Analysis Of VAriance) with video quality type as the independent variable and QoP-LoQ as a dependent variable, showed that video quality type significantly impacts user QoP-LoQ {F(1,8) = 7.706 p<0.001} (See 28a). Results supported the previous finding, which state that jitter and delay causes a significant drop in user QoP-LoQ

Interestingly, an ANOVA with video type (see Appendix B) as the independent variable and QoP-IA as a dependent variable, showed that video type was significantly affect the user QoP-LoQ {F(1,11) = 7.085 p<0.001} (See 28b). This is interesting as it suggests that video style (i.e. the way in which information is presented) is of significantly important to the user-perspective of multimedia video quality at the network-level, and therefore should be considered when defining multimedia quality. It also supports the manipulation of information style as a means of improving user QoP-LoQ at the network-level.

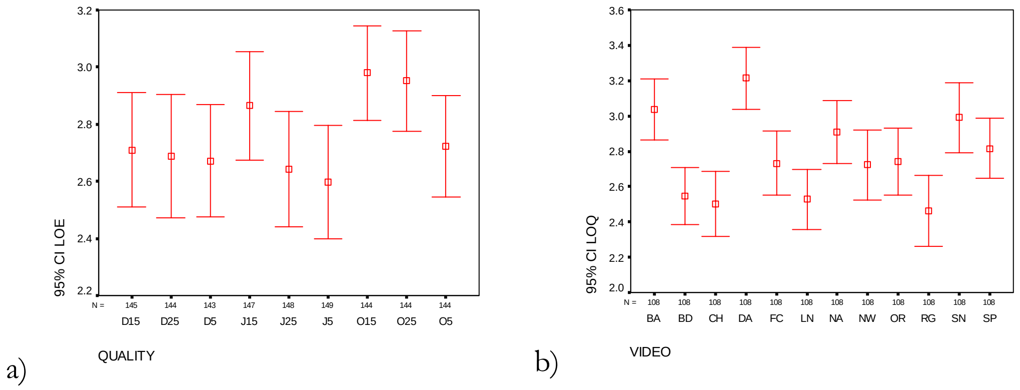


Figure 4.8: Impact of a) Quality type and b) Video type user QoP-LoQ {Mean and St. Dev.}.

4.4.3. Impact of Jitter and Delay on QoP-LoE

A MANOVA (Multiple ANalysis Of VAriance) test, with video variation type and video quality type as the independent variables and QoP-LoE (Level of Enjoyment) as a dependent variable, showed QoP-LoE to be significantly impacted both by variation type {F(1,2) = 3.954 p=0.019} and quality type {F(1,8) = 2.221 p=0.024} (see Figure 4.9, Figure 4.10a). Moreover Post hoc tests show important differences between the control and delay {p=0.019}, and control and jitter {p=0.037} videos, highlighting that both QoP-S factors (perception of video quality and user enjoyment) are significantly impacted by network-level quality parameter variation.

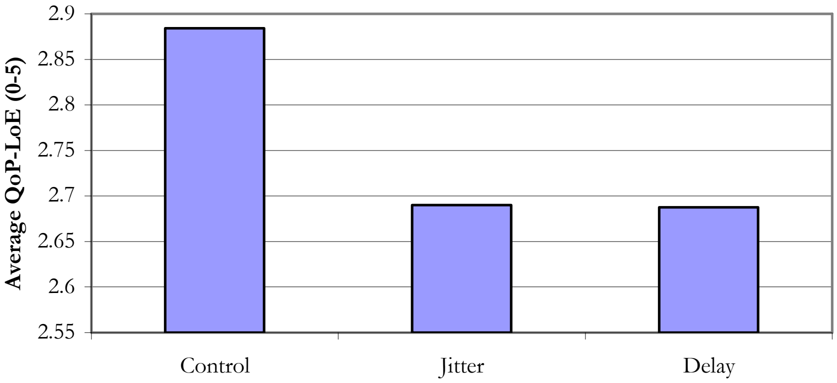


Figure 4.9: Impact of simulated network-level quality parameter variation on user QoP-LoE.

An ANOVA test with video type as the independent variable and QoP-LoE as a dependent variable, showed that video type significantly impacts user enjoyment {F(1,11) =8.322, p<0.001} (see Figure 4.10b). Importantly, results show that the type of video being presented is more significant to a user's overall QoP (i.e. both the level of information assimilated and user satisfaction) than either variation in presentation frame rate, or the introduction of error (jitter / delay).

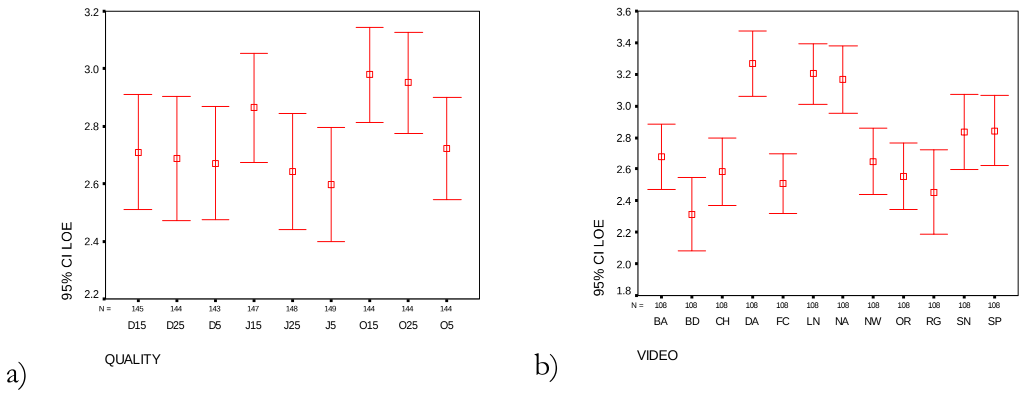


Figure 4.10: Impact of a) Quality type and b) Video type on user QoP-LoE {Mean and St. Dev.}.

4.4.4. Impact of Jitter and Delay on Eye-Gaze

4.4.4.1. Median Statistical Eye-gaze Analysis

To allow statistical correlation of eye-position between frame rates (5, 15 and 25 fps) and quality variation groups, over the duration of each video clip (between 640-1128 frames when shown at 25 fps playback), three coordinate points were required for each eye tracker sample, with each sample point relating to a specific frame rate group (5, 15 and 25 fps). Eye tracking samples (25 Htz) correspond to the maximum frame rate used in the experimental material, thus facilitating comparison between eye-tracking data and video frames. As we are not aware of any previous eye-tracking data analysis that uses statistical comparison across multiple video frames, there was no known precedent for summarising multiple participant eye-tracking data in this way. Accordingly, to avoid inclusion of extreme outlying points whilst removing unwanted data, such as error coordinates as a result of participant blinking, our study uses - for each eye-tracking sample, for each experimental data set (5, 15, 25 fps) - the median x and y coordinate values of participant eye-gaze.

Although a median value is not ideal, especially if multiple regions of interest exist within a frame, we considered it to be least prone to error values, yet still facilitate statistical analysis. By mapping these x and y median coordinate values in time we were able to calculate the median eye-path through each multimedia video clip, which we called the *video eye-path*, for all video clips shown at each of the video quality type (see Table 4.2). The example (see Figure 4.11) shows the control x coordinate value for the space video clip. The space clip shows a dynamically changing video based around a gun battle (see Appendix B). Although, in this specific example, eye fixations tends to return to position 5000 (the centre of the screen), this is not always the case. We can therefore assume that this trend is clip-dependent. Captured eye tracking data contains four dimensions of information: the x coordinate, the y coordinate, the distribution of samples in a specific screen area and time. Mapped median values represent two of the four possible data dimensions (a single coordinate value and time). Mapped median values reduce analysis complexity and facilitate statistical analysis.

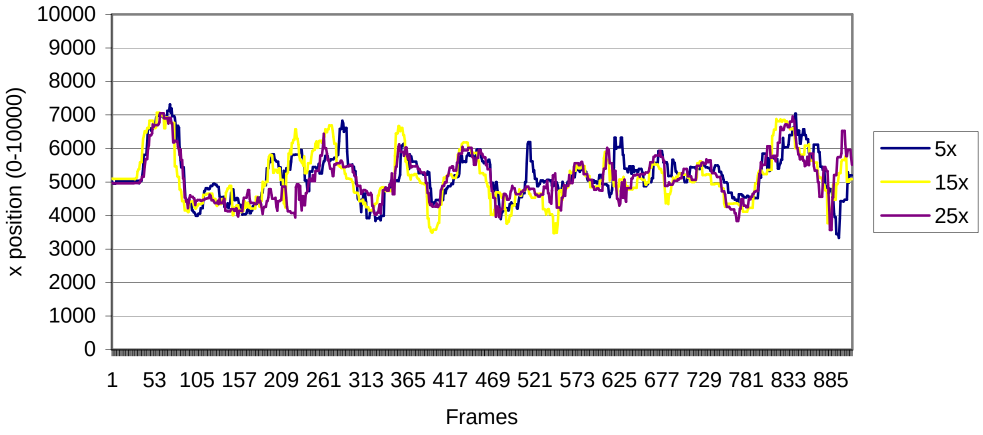


Figure 4.11: Space Action Movie x-coordinate video eye-path.

Statistical correlations were subsequently performed (Kendall's tau-b and Spearmans 2–tailed nonparametric tests) between median coordinate values, for eye-tracking samples of 5, 15 and 25 fps (i.e. 5fps compared to 15 fps, 5 fps compared to 25 fps, and 15 fps compared to 25 fps). This comparison was done for all of the 12 multimedia video clips used in our experiment. In addition, for each of the video clips, comparison was also made between control, delay and jitter video groups, e.g. O5X (original video at 5fps {*x coordinates*}) compared to D5X (delay videos at 5fps {*x coordinates*}). These tests were used to establish whether varied frame-rate or network-level quality parameter variation (delay and jitter) statistically impacted video eye-paths, i.e. that similar median trends of eye movement occur for groups of people when shown the same video content at different frame rates or with different network-level quality parameter variation.

Control: All control correlation tests showed a correlation value of p<0.001 between the video eye-paths across the different frame rates. This result shows that, for median coordinate values mapped across time, eye movement significantly correlates independent of the underlying video frame rate. With such strong correlation between participants, and the fact that strong correlation exists for all of the diverse multimedia video clips, we can conclude that frame rate does not significantly impact median control video eye-path.

Delay: All delay correlation tests showed correlation value of p<0.05 between the video eye-paths across the different frame rates. Interestingly, delay videos were involved in 10 of the 26 non-correlations that exist between different error groups (e.g. C25X and D15X), i.e. that the presence of delay in experimental videos causes slight variation in user eye-path, when compared to control and jitter videos. These delay non-correlations were only identified for BD (Band), CH (Chorus), DA (animation) videos, which suggests that the presence of delay is only significant for certain video content.

Jitter: Although the majority of jitter correlation showed a significant correlation (p<0.05) across frame rate variations, there was one noticeable exception - the band video clip (15fps / 25fps) {r(887) = 0.58, p=0.85}. In addition, jitter videos were involved in 22 of the 26 non-correlations in user video eye-path, which exists between different error groups (e.g. C5X and J5X), i.e. that the presence of jitter in experimental videos causes considerable variation in user eye-path, when compared to either control and delay videos. Interestingly, jitter non-correlations were also only identified for certain videos: BD (Band), CH (Chorus), DA (animation) and FC (Weather forecast) videos, which suggest that the use of jitter has more of an affect on these videos.

Results show that although the majority of video eye-paths correlate, addition of delay and jitter increases disparity in user video eye-paths, with jitter having more of an affect than that of delay. Interestingly, this disparity only happens for four out of twelve videos (BD, CH, DA, FC). Although this is most probably due to the existence of multiple Regions of Interest (RoI), this conclusion can not be made at this time.

4.4.4.2. Fixation Maps

To better understand the impact of jitter and delay on video eye-path we implemented fixation maps for control, jitter and delay eye-tracking data, for each video frame of the experimental videos.

Fixation maps were first introduced by Wooding [WOO02], who conducted the world's largest eye-tracking experiment, in a room of the National Gallery (LONDON), over the winter of 2000 / 2001, as part of the millennium exhibition. Over 3 months 5,638 participants had their eye movements successfully recorded, whilst viewing digitised images of paintings from the National Gallery collection. The quantity of the resultant data, at that time, was unprecedented and presented considerable problems for both understanding results as well as communication of results back to the public. Wooding proposed the use of fixation map, which is a novel method for manipulating and representing large amounts of eye tracking data. Fixation maps are better when considered as a terrain or a landscape (see Figs. 32b, Figure 4.12c and Figure 4.12d). Since fixation maps are produced from eye-tracking data, the colour at any point in the image represents the areas of an image/video frame that possess the greatest number of user fixations (0- no interest, 255- maximum interest), i.e. white areas represent the regions of interest (See Appendix F). Fixation maps, not only allowed regions of interest areas to be mapped in the form of an image, but also allow the difference between two data sets to be determined (See Appendix F). If control, jitter and delay fixation maps are produced for all video frames in experimental video material (approximately 120,000 in total), then the difference between the fixation maps for a specific video frame represents the difference between participant regions of interest, as a result of error type (see Figure 4.12e and Figure 4.12f). By analysing the average pixel value for consecutive difference fixation maps, we can identify specific frames where a higher level of user video eye-path variation exists, i.e.: a relative increase in average pixel value, as a result of delay and jitter (see Figure 4.13 and Appendix G).

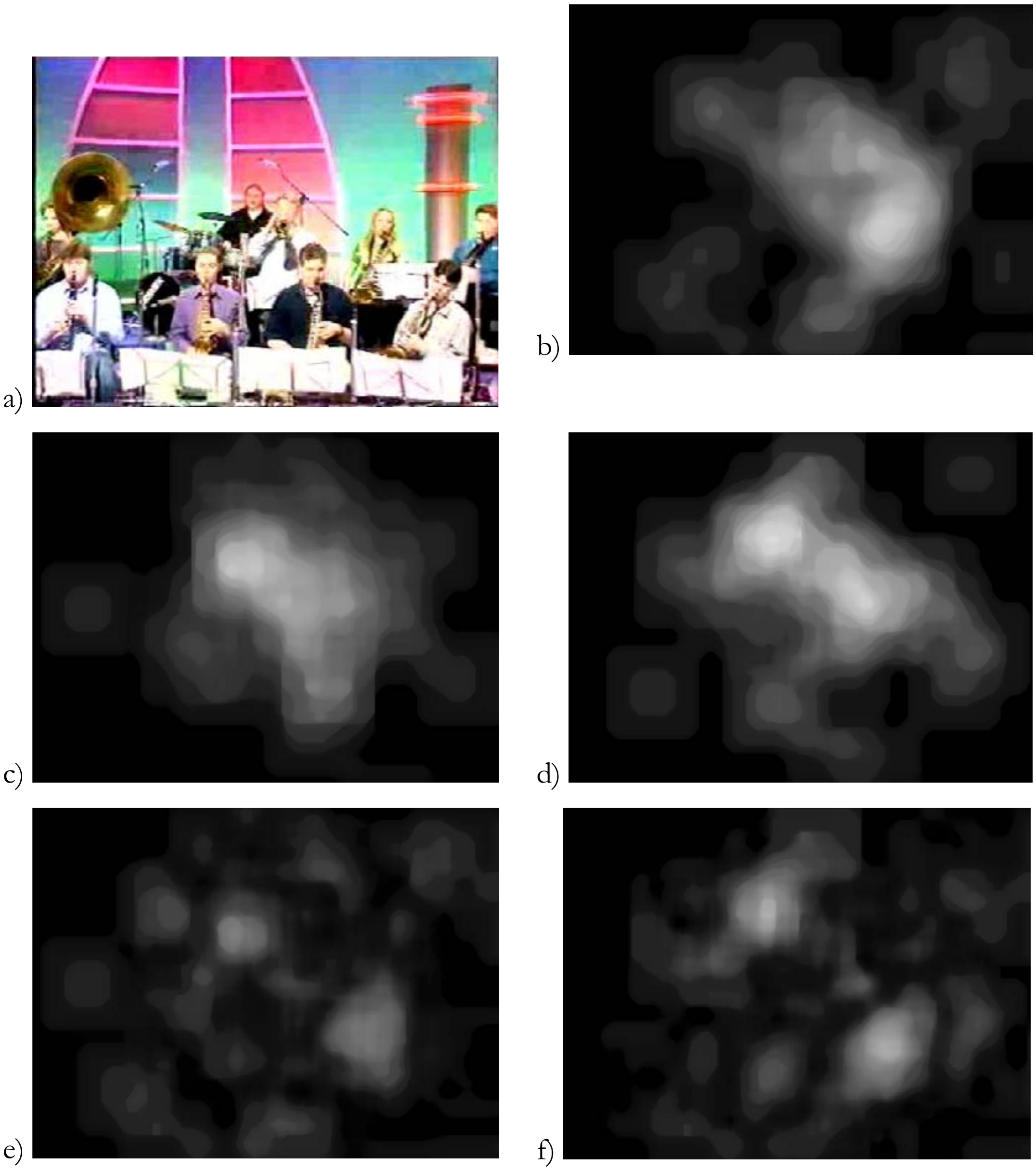
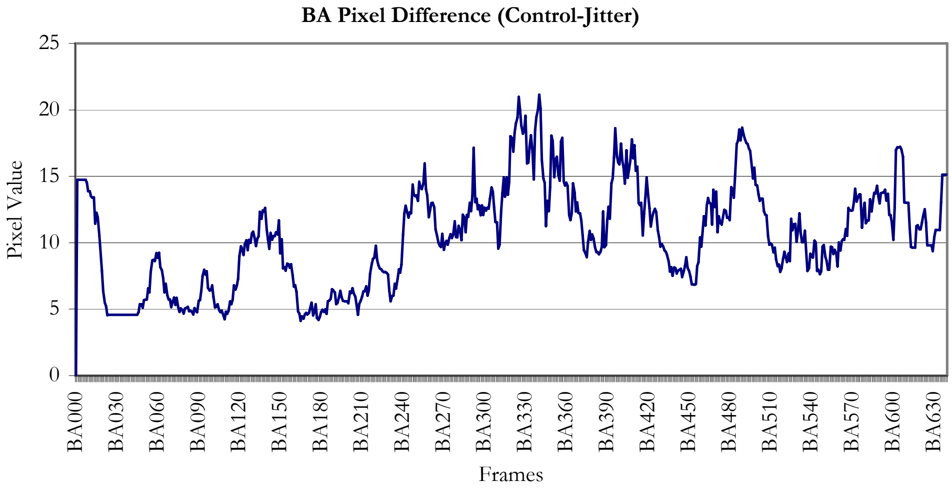


Figure 4.12: a) original frame; b) control fixation map; c) delay fixation map; d) jitter fixation map; e) pixel difference between control and delay RoI areas; f) pixel difference between control and jitter RoI areas.

Analysis suggests that high levels of disparity in user eye-path occur for two reasons: i) when no single / obvious point of focus exists, e.g. frames 440-490 in the NA video clip, which represents the introduction of the second pop-fact (see Appendix B), thus causing a conflict of user attention; or ii) when the point of attention changes dramatically, e.g. in the BD clip, where a scene is only shown for a matter of a few seconds, which does not provide the user with enough time to identify regions of interest and subsequently adapt his/her eye-position. Low variation in user-eye path occurs when a single point of attention exists, e.g. frames 150-350 in the CH clip, which highlights a solo singer or in the NW clip, frames

150-650, which shows the newsreader after the information concerning the channel and newsreader's name has been removed (see Appendix G). Figure 4.13 illustrates these ideas and shows the impact of i) multiple stimulus and ii) rapidly moving single stimuli, on user video eye-path in the BA video.



Frame Numbers	Video Content	Comment
0 - 37	i) GMTV logo ii) Moving ball	Multiple points – one moving.
38 – 59	Women speaking	Single (obvious) point of focus.
59 – 100	i) Man in the bath (talking) ii) Cream cleaner	Multiple points of focus.
101 - 120	Man in the bath (talking)	Single point of focus.
121 - 180	i) Man in the bath (talking) ii) Moving arms	Moving multiple points of focus.
181 - 240	Moving sponge duck	Rapidly changing single point of focus.
241 - 270	i) Man (talking) ii) holding bottle of cleaner.	Multiple points of focus.
271 – 310	i) moving cleaner bottle ii) bath fittings / cleaner liquid	Multiple points of focus.
311 - 370	i) bath fittings ii) moving sponge / finger	Multiple points of focus.
371– 400 / 401 - 450	i) Man (talking) ii) Full screen point of focus	Multiple points of focus.
451 - 510	i) Couple talking in background ii) Bathroom suite	Multiple points of focus.
511 - 570	Women talking	Single point of focus.
571 – 640	i) Cleaning bottle ii) Advert text	Multiple points of focus.

Figure 4.13: BA video Pixel Difference, between control and jitter fixation maps, plus critical analysis.

It is important to note that dynamic video content does not negatively impact user region of interest, as long as a small single point of focus exists, e.g. in the rugby clip, it is only after the score (i.e. after the ball has been removed), that the highest variation in video eye-path exists. If a full screen or large single stimuli exists then variation in user eye-path occurs, e.g. the man wiping the bath with the sponge (see Figure 4.13).

26 non-correlations of video eye-path were found between different error groups, which suggests that the presence of delay and jitter introduces variation in user eye-path. Interestingly, these non-correlations only existed for BD (Band), CH (Chorus), DA (Animation) and FC (Weather forecast) videos, which, for the majority of the video, do not have a single / obvious point of focus. Instead multiple conflicting points of focus exist in BD, CH, DA and FC, which ultimately causes variation in user eye-path, as a result of delay and jitter. Moreover, the point of focus in BD and DA video changes dramatically as a result of fast scene changes or multi person dialogue, i.e. does not ensure smooth pursuit eye movement, thus resulting in problems identifying and tracking important points-of-focus / regions-of-interest.

4.5. Conclusion

In this chapter, we measured the impact of delay and jitter on user perception of multimedia quality. Results showed no difference in the level of information assimilation, as a result of video variation type (i.e. control, jitter and delay), which demonstrates that delay and jitter do not negatively impact information assimilation; an important result for distributed educational applications, as it shows that QoS degradation does not impact the users ultimate understanding of the video content at the network-level. The type of video clip, however, was found to significant impact user QoP-IA, thus supporting the use of video material (see chapter 3).

Video variation type was found to significantly affect both user QoP-LoQ and QoP-LoE, which suggests that not only can a user distinguish between a video presentation with and without error, but the presence of error impacts the user's level of enjoyment. These are important findings, especially in band-width constrained environments, as it means that the user's QoP-S is impacted by network-level QoS variation. If distributed multimedia is used for entertainment purposes, where user QoP-S is critically important, then it is critical that network QoS variation is minimised in order to minimise impact on user QoP-S. Moreover, findings support Procter et al. [PRO99], who observed that degradation of network-level QoS has a greater influence on a subjects' uptake of emotive / affective content than on their uptake of factual content.

In this chapter, we incorporated eye-tracking with the QoP concept in order to facilitate continuous monitoring of attention and therefore provide a better understanding of the role that the user plays in the reception, analysis and finally the synthesis of multimedia data. Moreover, in this chapter we have introduced two eye tracking data analysis methods (median- and fixation map- analysis), which when considered together allows analysis of the 4-dimensions of eye-tracking data: i.e. x-coordinate values, y-coordinate values, the distribution of samples showing the importance of an image region, and time. Both methods, although not ideal, provide important information concerning user attention and video eye-path. Results show that variation in video eye path occurs when: i) no single / obvious point of focus exists; or ii) when the point of attention changes dramatically. Further research, determining other ways of considering the 4-dimensions of eye-tracking data, would benefit eye-tracking analysis, yet lies outside the scope of our study.

In conclusion, in this chapter we measured the impact of network-level parameter variation (delay and jitter) on the user's perception of multimedia quality. In the next chapter we aim to measure media-level technical- and user-perspective parameter variation, thus fulfilling research objective 2. Subsequently, we shall continue to consider user attention, by developing a novel frame rate-based attentive display, which manipulated video using region-of-interest determined from both video content- and user- dependent output video.

CHAPTER 5

Media-Level Quality Parameter Variation

Our work aims to extensively consider the user's perceptive regarding multimedia quality, by adapting relevant technical- and user-dependent parameters at all quality abstractions: network-level (technical-perspective), media-level (both technical- and user-perspectives) and content-level (both technical- and user-perspectives). In this chapter, we measure the impact of media-level quality parameter variation on user perception of multimedia quality by implementing an *attentive display*, which facilitates media-level variation as a result of both eye tracking- (User-Perspective) and video content-based (Technical-Perspective) data. Eye-tracking data will be used at the media-level to manipulate video content, yet no monitoring of user eye-gaze location will be made at the media-level.

5.1. Introduction

Visual information is computationally intense, however due to improving rendering hardware, high QoS digital video is commonplace. With increased screen sizes and screen resolutions, the user has a growing expectation of what defines high quality video [BAR96]. Although increased provision is possible with High Definition TeleVision (HDTV) and DVD technologies, growing user expectation places considerable pressure on multimedia video when transmitted over bandwidth constrained environments, as limitations in bandwidth also restricts the quality variation that can be made at the media-level.

Interestingly, although electronic displays and computing devices are inextricably linked [HOF78], most of the resources used to produce large, high resolution displays are wasted, as the user never looks at the whole screen at one point in time. As described in Chapters 2 and 3, high acuity colour vision relies on the uneven distribution of cone receptors on the retina. Accordingly, ocular physiology limits the range of high acuity to approximately 2° of the visual field, which is equivalent to approximately the width of your thumbnail at arms length or 2cm at a typical reading distance of 30cm [BAU03]. If the *Human Visual System* (HVS) can only process detailed information within an area at the centre of vision, with rapid acuity drop-off in peripheral areas [MAC70], the position of user attention, if effectively monitored and/or predicted, can be used to manage the non-uniform allocation of bandwidth. Displays that facilitate this relationship between bandwidth and user point of gaze are known as attentive displays [BAU03]. Two main approaches have been developed to implement attentive displays: *Gaze-contingent Display* (GCD) and *Region of Interest Display* (RoID) systems. Attentive GCDs select the user point of focus by actively tracking the viewer's eyes in real time and maintaining a high level of detail at the point of gaze. On the other hand, RoIDs define user *Regions of Interest* (RoI) coordinates, by either analysing previously obtained eye-tracking data or by analysing the characteristics of the video content, and adapting the displayed video quality such that resource allocation is biased to RoI areas.

Although variation in multimedia systems has been widely employed since their inception, such variation has traditionally focused on areas such as media streaming [HES99; KRA03; MAH03], personalisation [DOG02; HAR03] and education [BLO03; MAY97]. Variation via media streaming concerns itself with the fine tuning of technical parameters such as bandwidth and transmission rates, or with the construction of appropriately tailored transmission protocols. Personalisation deals with the effective variation of multimedia content by incorporating user preference in the choice of subject matter and media presentation settings, whilst education concerns itself with the variation of media to support information transfer. Personalisation and education consider the user-perspective, however the former almost completely ignores the user understanding of video content and the latter rarely implements manipulation of media-content based on physiological data.

In this chapter we measure the impact of media-level quality parameter variation, from both technical- and user-perspectives, by considering the perceptual impact of adapting RoIDs manipulated by either empirically obtained eye-tracking data or computationally defined RoI data. Accordingly, the structure of this chapter is as follows: Section 5.2 introduces user attentive processes and current attentive displays, making specific reference to the perceptual impact on user perception. In section 5.3 we provide specific details concerning the experimental approach used in this chapter. Results are presented and discussed in section 5.4, which is followed, in section 5.5, by our conclusions.

5.2. Attentive Processes and Displays

5.2.1. Attentive Processes

If cones were distributed evenly across the retina, their average distance apart would be relatively large, and the ability to detect fine spatial patterns (acuity) would be relatively poor. Cones are therefore concentrated in the centre of the retina, in a circular area called *macula lutea*. Within this area, there is a depression called the fovea, which consists almost entirely of cones, and it is through this area of high acuity, extending over just 2° of the visual field, that humans make their detailed observations of the world. High acuity colour vision therefore relies on cone receptors located in this area of the retina. Movement of the eye, head and body are used to bring a Region of Interest (RoI) into the visual path at the centre of the fovea. This movement between items within the stationary field, the eye field and the head field is determined by visual attention [SAN70].

The process of visual attention can be broken into two sequential stages: the *pre-attentive* stage and the *selection* stage [SAL66; TRE86]. In the pre-attentive stage, information is processed from the whole visual field in parallel. It is the pre-attentive stage that determines RoI within the visual field (defining important visual cues) and based on this pre-attentive mapping the selection stage performs high-level serial processes, which are dependent on high-level cognitive search criteria (see Table 5.2). When objects pass from the pre-attentive stage to the selection stage, they are considered to be selected [YAR67].

5.2.1.1. Pre-attentive stage

We do not see the world as a collection of colours, edges and blobs. Instead we organise the world into defined surfaces and objects. This is because the pre-attentive stage of vision subconsciously defines objects from visual primitives, such as edges, orientation, colour and motion [SAL66].

Table 5.1: Low Level Factors Impacting RoI in the Pre-attentive Stage.

Low Level Factor	Description	Supporting Work Conducted By
Contrast	Regions that have a high level of contrast (the relation between intensity in colour) attract greater attention.	Findlay [FIN80], Senders SEN92, Yarbus [YAR67]
Size	The greater the size of an object, the more likely it is to attract attention.	Findlay [FIN80] Mordkoff and Yantis [MOR93]
Shape	Long and thin shapes are more likely to attract attention than round shapes.	Gale [GAL97], Lappin [LAP00], Sender [SEN92]
Colour	Certain colours are more sensitive to the eye. Colour also has social meaning / importance.	Cole [COH90], Mordkoff and Yantis [MOR93], Niebur and Koch [NIE97]
Motion	Motion is the strongest influence on visual attention.	Hillstrom and Yanis [HIL94]
Brightness	Bright areas attract a greater level of attention than dark areas.	Kruger et al. [KRL04]
Line Ends	Distinctive edges, separating definite regions, attract more user attention.	Hubel and Wiesel [HUB70] Yarbus [YAR67]
Orientation	Changing the orientation of an area stimulates visual attention.	Kruger et al. [KRW04]

The pre-attentive stage of vision operates in parallel across the entire visual field and has no measured capacity limitations. Learnt visual schemas are used to define how visual primitives are grouped into ‘chunks’ and how these ‘chunks’ are then perceived as objects. It is in the pre-attentive stage, using low-level factors (see Table 5.1), that humans determine objects within the visual field.

5.2.1.2. Selection Stage

Based on the pre-attentive mapping, the selection stage determines contextually significant areas of the visual field, dependent on cognitive high-level criteria (see Table 5.2). When the pre-attentive and selection stages have determined the position of the target, the eye must be moved in such a way that the target object can be inspected with a higher acuity, by *foveating* the object, i.e. bring to the centre of the fovea, thus ensuring acuity. As mentioned in Chapter 3, the principal method for moving the eye position to a different part of the visual scene is through the use of *saccades*, which are sudden, rapid, ballistic movements of the eyes. During a saccadic movement the processing of the visual image is suppressed, therefore processing of the retinal scene occurs between saccadic periods, which are periods of time called fixations that last between 200 and 600ms.

Table 5.2: High Level Factors Contributing to Region of Interest Formation in the Selection Stage.

High Level Factor	Description	Supporting Work Conducted By
Location	25% of viewers concentrate on or around the centre of the screen.	Elias et al. [ELI84]
Foreground / Background	Foreground objects are considered more contextual relevant than background objects.	Cole et al. [COH90] Yarbus [YAR67]
People	Eyes, faces, mouths and hands are areas of significant importance to human gaze.	Gale [GAL97], Senders [SEN92], Yarbus [YAR67]
Context	Eye movements can drastically change depending on the instruction given whilst watching an image.	Gale [GAL97], Yarbus [YAR67]

The eye naturally selects / fixates on, areas that are likely to be most informative [KAU69]. As well as the definition of 'informative' being contextually dependent on high-level processes (see Table 5.2), four distinct looking states have been defined that summarise the cognitive state of the user [KAH73]:

- *Spontaneous looking* - when a subject is not actively looking for, or thinking about, any specific object, e.g. looking at a picture without task or instruction.

- *Task-relevant looking* - when a subject is performing a specific task, such as reading text or inspecting a picture in context of specific instructions.

- *Orientation of thought looking* - eye movements of this kind represent a general orientation towards the object of thought, e.g. when a subject thinks of a object within their visual field, he /she will feel a tendency to look at that object.

- *Intentional manipulatory looking* - when subjects consciously control their direction of looking to provide output to a visual guided control system, e.g. an eye-tracker controlled graphical user interface. Eye-tracking equipment measures either the saccadic movement of the eye or the point of fixation and are accordingly categorised as being either fixation or saccadic pickers [JUS76]. By measuring user eye-movement, or modelling user attention, it is possible to manipulate multimedia video content, thus providing a higher level of video quality around the user's point of gaze. Such displays are termed attentive displays.

5.2.2. Adaptive Attentive Displays

Stelmach et al. [STT91] showed that eye-movements during television viewing are not idiosyncratic to a specific viewer, but instead that the direction of gaze is highly correlated amongst viewers. This view is supported by the findings of chapter 4 [GULL04] and supports the use of attentive displays, especially in bandwidth-constrained environments, with the aim of minimising bandwidth requirements and negative user perception of multimedia quality. Two main approaches have been developed to implement attentive displays: Gaze-contingent display (GCD) and Region of Interest Display (RoID) systems. Attentive GCDs select the RoI by actively tracking the viewer's eyes in real time and maintaining a high level of detail at the point of gaze. On the other hand, RoIDs use RoI coordinates, obtained from either prior eye-tracking data or from analysing of characteristics of the video content, to adapt the video being displayed such that resource allocation is biased towards RoI areas. GCD systems necessitate an eye tracker device with high sample rates, with corresponding computational, technical and cost implications, to ensure appropriate refresh rates (4-15ms depending on window size [LOS00; REI02]). GCDs only support a single user, however they do facilitate real time attention-based rendering. RoIDs can be configured to accommodate multiple users and can facilitate distributed bandwidth savings, as video can be coded, reducing bandwidth requirements in non-RoI locations, prior to transmission. Moreover, RoIDs boast a series of advantages (over GCDs): RoIDs can be achieved at a fraction of the cost of GCDs, can be used by multiple users, do not require the user to possess specialised or additional hardware, are easily integrated with current display systems and can reduce transmission bandwidth requirements. These make RoIDs commercially attractive, especially in bandwidth constrained environments.

Early attempts at variable quality attentive displays suffered from limited display sizes, noticeable quality edges and limited control of resolution [BAR96; JUD89; KOR96; SIL93; WEI90]. However, increased screen and resolution sizes, as well as the falling cost of eye-tracking equipment, have all led to increased interest in attentive display research. Current attentive display techniques were first introduced in [MCC75, SAI79] and are used in a wide range or applications, including: reading, perception of image and video scenes, virtual reality, computer game animation, art creation and analysis as well as visual search studies [BAU03; PAR02, WOO02].

A number of recent studies, which looked at the perceptual impact of using attentive displays, however made mixed conclusions concerning their ultimate usability. Accordingly, Reingold and Loschky found that when they adapted a high-resolution window at the point-of-gaze and degraded resolution in peripheral areas, participants had longer initial saccadic latencies in peripheral areas (the time taken to identify a visual target), than when a low resolution was uniformly displayed across the whole display window [REI00]. Loschky and McConkie found, in support of earlier studies [SHI89; WAT96], that if degradation is increased in peripheral areas, the size of the adapted high-resolution window at the point of gaze also needs to be increased, if the users level of task performance is to be maintained [LOS00].

These findings are important, as the use of high-resolution/quality regions is central to attention-based displays. Such results place into doubt the real effectiveness of windowed gaze-contingent multi-resolution displays, especially in applications such as pilot simulations, where any delay to peripheral stimuli or reduction in performance may have severe consequences to user information assimilation.

Reingold and Loschky suggested that reduced reactions may be due to participants being distracted by sudden boundary edges (a change in visual resolution). Consequently, Reingold and Loschky compared sharp and blended resolution boundary conditions, to identify whether increased saccadic delays were due to boundaries lines [REI01]. Reingold and Loschky used three conditions: i) the no-window condition (where the entire image is blurred - of lower quality); ii) a 12° window, which implements no blending; and finally iii) a 12° window with a 3° wide region of where resolution is blended. Results showed that the no-window condition produced a shorter mean initial saccade, thus supporting the work of Czerwinski et al. [CZE02] who stated that a wider field of view increases performance in productivity. Interestingly, Reingold and Loschky found no difference in a user's ability to identify visual errors, as a result of the window edge being either sharp or softened. This is in distinct conflict with previous work [BAL81; TUR84], which indicated that the blending regions were vital to the perceptual quality of the display.

The arguments of [LOS00] and [REI00] place considerable doubt on whether attentive displays provide any perceptual advantage, especially in low-bandwidth environments. Interestingly, Osberger et al. used a technique for controlling adaptive quantisation processes in an MPEG encoder, based on frame-based Importance Maps (IMs) [OSB98b], and provided diametrically opposed conclusions concerning the usability of attentive displays. IMs, similar in appearance to Fixation Maps, are produced using segmented images, which are analysed using five attention factors, namely: contrast, size, shape, location and background importance [OSB98a]. Lower quantisation was subsequently assigned to visually important regions, whilst areas that are classified as being of low visual importance were more harshly quantised. Osberger et al. tested their method on a wide variety of images with results showing that IM based adaptation significantly correlates with human perception of visually important regions.

Different implementation approaches (GCD or ROID systems), as well as use of different adaptation data (eye-based and content-dependent RoI data), appear to have a significant impact on the users reaction and ability to notice errors. Interestingly, no studies to the best of our knowledge have considered the impact that either attentive GCDs or ROIDs have on user understanding of multimedia content, and more generally on the user-perspective regarding multimedia quality (i.e. QoP). Accordingly, we measure the impact of media-level quality parameter variation on user perception of multimedia quality by implementing a *Region-of-Interest Display* (RoID) that incorporates frame-rate as a result of both eye tracking- (user-perspective) and video content-based (technical-perspective) data. We chose RoI-based frame rate as the main QoS parameter of interest within this chapter, because it is the main factor affecting multimedia bandwidth requirements, and is of primordial importance, due to its scarcity, in distributed multimedia environments.

5.3. Obtaining RoI Data

To implement a RoID system, both video content-dependent and eye-based RoI data was required. Moreover, consistent RoI scripts and adaptation software were required. Accordingly, in this section we describe how video content-dependent and eye-based data was obtained, defined and manipulated.

5.3.1. Video Content-Dependent RoI Data

In our study, video content-dependent RoI data was obtained through automated analysis of video content. This was a two stage process, which computed *primitive images* and consequently extracted RoI information. Accordingly, the primitive images considered in our study were based on the visual primitives of edges, movement and colour contrast, identified in the literature [COH90; HILL94; HUB70; MOR93; NIE97; YAR67] as being perceptually relevant:

Colour contrast images - A 640x480 pixel image was extracted for each video frame (see Figure 5.2a), for all of the twelve video clips described in Chapter 3 (see Appendix B). These colour images were subsequently used to calculate image areas with a high level of colour contrast.

Edge images - Edges characterise boundaries and are therefore fundamentally important in image processing. Most edge detection methods work on the assumption that this is a very steep gradient in the image, accordingly by using a weighted mask it is possible to detect edges across a number of pixel values. The simplest gradient operator is the Roberts Cross operator (see Figure 5.1), which uses the diagonal directions to calculate the gradient vector. The Robert Cross operator was applied to each of the extracted frames producing a black and white 640x480 pixel image, clearly displaying edge regions (see Figure 5.2b).

+1	0
0	-1

Gx

0	+1
-1	0

Gy

Figure 5.1: Roberts Cross convolution kernels [FIS03].

Movement images - The pixel difference between frame N and N+1 determined the quantity and location of movement in subsequent video frames. Software, see Appendix F, was developed to identify the pixel difference between two 640x480 images, with the output displaying areas of movement (see Figure 5.2c).

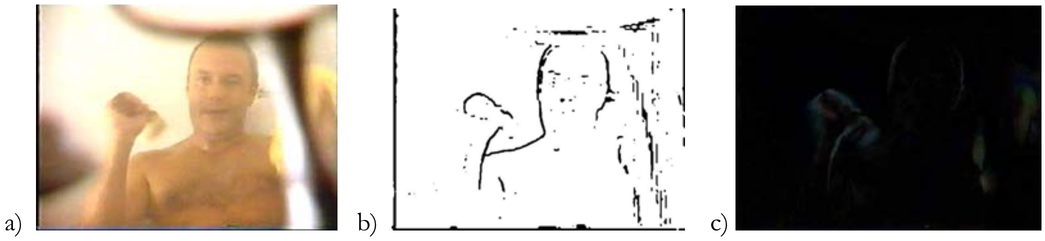


Figure 5.2: a) Original video frame, b) Edge detection in video frame, c) Motion detection in video frame.

Software was developed (see Appendix H) to calculate the distribution of the RGB pixel values for each colour, edge and movement image. This allowed colour, edge and movement mean pixel and standard deviation values to be determined for each image for all of the experimental videos (see Appendix I, Appendix J and Appendix K respectively). By combining image data, a mean pixel and standard deviation value was calculated for each of the twelve video clips. Important regions of colour, edges and movement were identified, assuming that:

- for colour, an abnormal distribution of colour occurs due to either an area of contrast (sharp difference in colour) or an abnormal colour value (see Appendix I).

- for edges, an abnormal average pixel value suggests a greater level of black lines, i.e. edges (see Appendix J).

- for movement, a higher than average pixel value suggests a greater variation level of pixel values between frame, i.e. movement (see Appendix K).

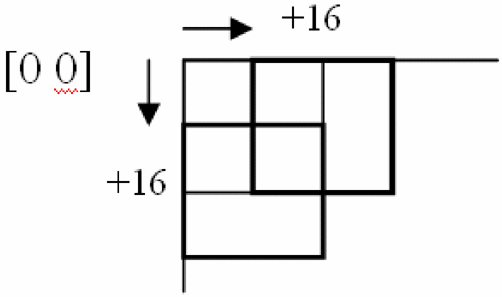


Figure 5.3: Overlapping pixel squares.

To calculate significant areas of the image, the colour, edge and movement images were split into overlapping 32x32 pixel squares (Figure 5.3). There were 300 squares used for each video frame, each representing a degree of the visual angle (both horizontally and vertically). A distribution of the RGB pixel values in each square was made, allowing mean pixel and standard deviation values to be determined for each 32x32 pixel square. A 32x32 pixel square was considered as being important if the mean pixel value (+/- standard deviation) was greater or less than either the mean pixel value for a specific frame (+/- frame pixel standard deviation), or the mean pixel value for the specific video (+/- video pixel standard deviation). As considerable variation in colour, level of edges and movement is possible in both a particular frame, as well as across consecutive frames, it was considered equally important to include both conditions (see Appendix I, Appendix J and Appendix K).

A 16-pixel shift between representing half a degree of the visual angle ensured that the majority of the image is covered by four separate comparisons, however the image edges are only covered by two comparisons and the corner 16x16 pixels are only covered by one.

5.3.2. Eye-based Data

To obtain eye-based RoI data, whilst ensuring that participants 'type of looking' was consistent with perceptual experiments, it was vital that the processes used to extract RoI data were consistent with the processes to be used in perceptual experiments. Accordingly, it made sense to actually use control eye-tracking data extracted in Chapter 4.

5.3.3. Producing RoI Scripts

Areas deemed as having important content were combined in a single RoI script for each multimedia video clip (see Appendix L). These scripts were subsequently used to adapt RoIDs. Eye-based and content-dependent RoI for a specific frame (see Figure 5.2a) can be seen in Figure 5.4a and Figure 5.4b respectively. RoI scripts, detailed in Appendix L, provide a considerably flexible and effective form of RoI data definition and may be incorporated by other attentive display systems.

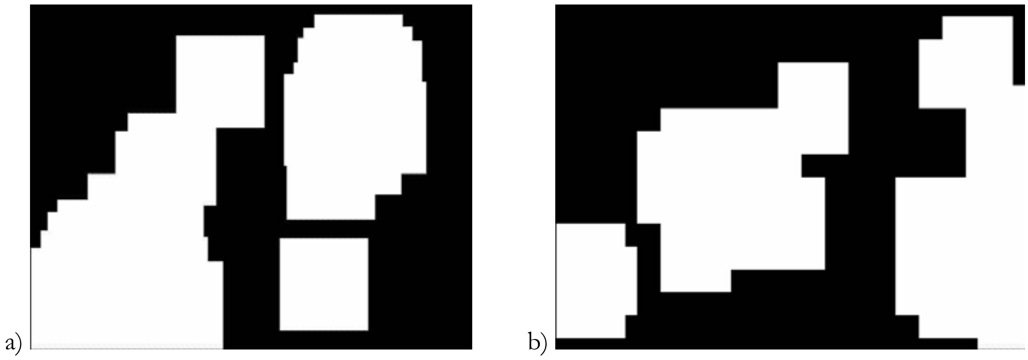


Figure 5.4: a) Eye-based RoI areas, b) Content-dependent RoI areas {white areas represent RoI}.

5.3.4. Producing Eye-Based and Content-Dependent RoIDs

To create RoIDs, we needed to produce video that had an adaptive non-uniform distribution of resource allocation. To achieve this we used eye- and content-dependent RoI scripts to adapt the frame rate in particular regions of the screen. Thus RoI areas, herewith referred to as foreground areas, were refreshed at a relatively higher frame rate than that of the non-RoI areas (background areas). Considerable effort was taken to make sure that each RoI foreground square covered at least 4° of the visual field (+/- 2° around the point of gaze), thus ensuring that the high acuity area of the fovea was contained within foreground area.

Specialised software (see Appendix M) was developed, using the Java Media Framework, which takes the original video (at 25fps) and a RoI script (either containing eye-based or content-dependent RoI data) and, using a 5 frame count, produces a playable multi-frame rate RoI-based MPEG video that presents the foreground and background regions at different frame rate combinations (see Figure 5.5). At playback, this video can be considered as a RoID, as it incorporates user eye-based (user-perspective) or video content-dependent (technical-perspective) RoI data.

To identify how varied foreground and background frame rate impacts user perception, our study considered three possible foreground and background combinations. Accordingly, nine video quality variation will be considered as part of our experiment: control 25fps (c25), control 15fps (c15), control 5fps (c5), eye-based and content-dependent 25fps foreground / 15fps background video (e25_15, v25_15); eye- based and content- dependent 25fps foreground / 5fps background video (e25_5, v25_5) and, finally, eye- based and content- dependent 15fps foreground / 5 fps background video (e15_5, v5_5).

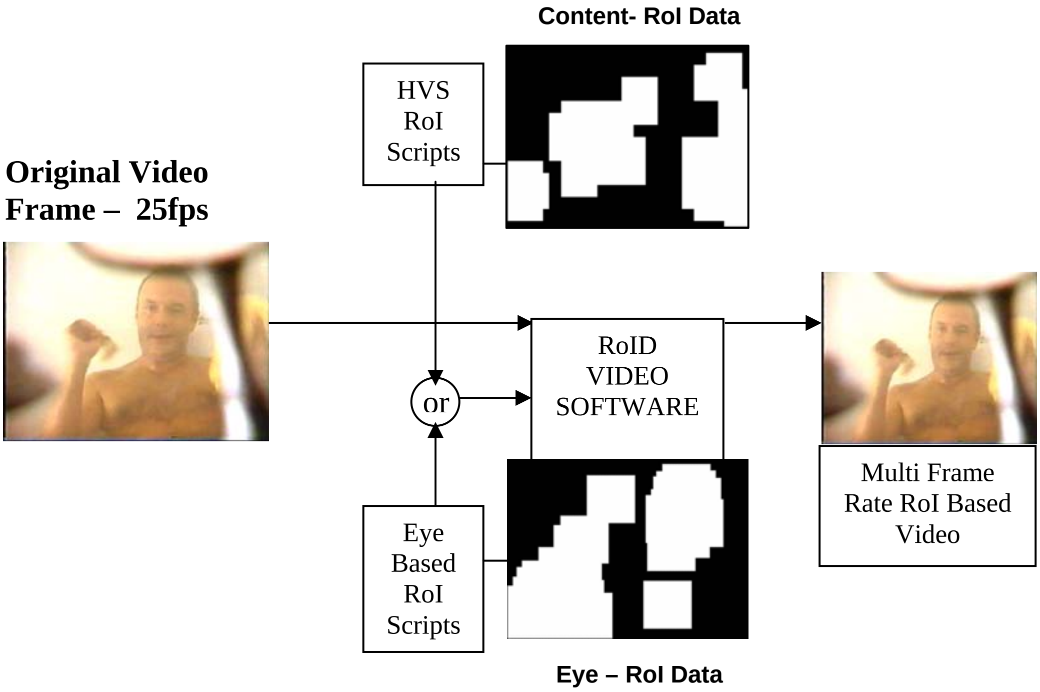


Figure 5.5: Process of adapting RoIDs from eye-based content-dependent data.

5.3.5. Experimental Variables

Three experimental variables were manipulated in this chapter: RoID presentation technique (i.e. control, eye-based and content-dependent data), multimedia video frame rate combinations, and multimedia content. Accordingly, both eye- and content-based RoID video was considered in our experiments. Moreover, three multimedia video frame rates: 5, 15 and 25 fps were used for each data quality variation type, using three foreground/background combinations: 25/15, 25/5 and 15/5 fps. As far as multimedia content is concerned, 12 video clips (see Appendix B) were considered in our experiments.

5.3.6. Perceptual Experiments

Perceptual experiments were carried out to measure the impact of media-level quality parameter variation on user perception of multimedia quality by implementing a *Region-of-Interest Display* (RoID) that incorporates media variation as a result of both eye tracking- (user-perspective) and video content-based (technical-perspective) data.

5.3.6.1. Participants

In our perceptual experiment a within-subjects design was used to ensure that participants view all nine video quality variation types (c25, c15, c5, e25_15, e25_5, e15_5, v25_15, v25_5, v5_5) across the 12 videos. Accordingly 54 participants, aged between 21 and 59, were evenly divided into nine experimental groups of six, with video quality shown in the order described in Table 5.3.

Table 5.3: Order of Video Quality Variations in Media-level Perceptual Experiments.

	Order 1	**Order 2**	**Order 3**	**Order 4**	**Order 5**	**Order 6**	**Order 7**	**Order 8**	**Order 9**
BA	C5	C15	C25	E25_15	E_25_5	E15_5	V25_15	V25_5	V15_5
BD	V15_5	C5	C15	C25	E25_15	E_25_5	E15_5	V25_15	V25_5
CH	V25_5	V15_5	C5	C15	C25	E25_15	E_25_5	E15_5	V25_15
DA	V25_15	V25_5	V15_5	C5	C15	C25	E25_15	E_25_5	E15_5
FC	E15_5	V25_15	V25_5	V15_5	C5	C15	C25	E25_15	E_25_5
LN	E_25_5	E15_5	V25_15	V25_5	V15_5	C5	C15	C25	E25_15
NA	E25_15	E_25_5	E15_5	V25_15	V25_5	V15_5	C5	C15	C25
NW	C25	E25_15	E_25_5	E15_5	V25_15	V25_5	V15_5	C5	C15
OR	C15	C25	E25_15	E_25_5	E15_5	V25_15	V25_5	V15_5	C5
RG	C5	C15	C25	E25_15	E_25_5	E15_5	V25_15	V25_5	V15_5
SN	V15_5	C5	C15	C25	E25_15	E_25_5	E15_5	V25_15	V25_5
SP	V25_5	V15_5	C5	C15	C25	E25_15	E_25_5	E15_5	V25_15

5.3.6.2. Perceptual Experiment - Set-up

To help simulate realistic distributed RoIDs viewing conditions, eye tracking was not used in this experiment. To ensure that experimental conditions remained consistent, the same experimental equipment was used for all participants. An HP mobile laptop AMD Athlon™ XP 2000+, with an inbuilt 15-inch LCD monitor and a ATI Radeon IGP 320M, was used to display manipulated video with a resolution of 640x480. Manipulated video was embedded in an Internet Explorer browser (see Figure 5.6), thus further simulating more realistic distributed RoIDs viewing conditions. All participants wore headphones, to ensure that a consistent audio level (70dB) was used for throughout all experiments.

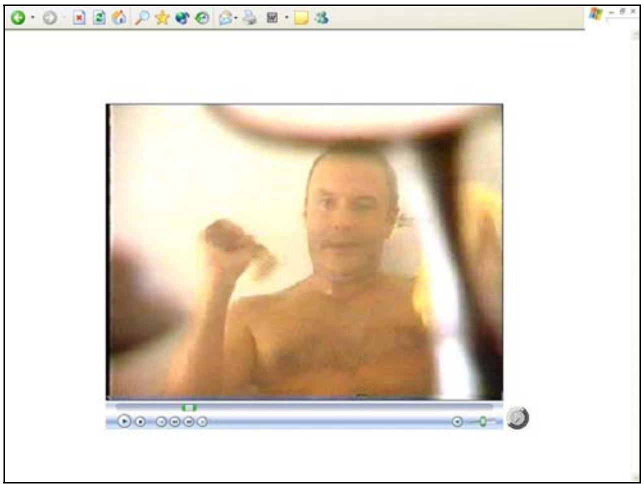


Figure 5.6: RoID video presented via embedded video in MS-Internet Explorer.

5.3.6.3. Experimental Process

Once the laptop and headphones were setup, an experimental process was used that is consistent with that defined in section 3.6. Appropriate experimental stimulus was loaded, with the video quality variation relating to one of the nine experimental groups, described in Table 5.3.

5.4. Results

5.4.1. Impact of RoID Quality on QoP-IA

QoP-IA was expressed as a percentage measure, which reflected the level of information assimilated from visualised multimedia content. An ANOVA test, with video quality variation (i.e. c25, e25_15, v25_5, etc.) as the independent variable and QoP-IA as the dependent variable, showed that both variation in RoID presentation technique and foreground/background frame rate combinations, do not have a significant effect on the user's level of information assimilation and understanding {F(1,8) = 1.321, p=0.230} (see Figure 5.7a). This results shows that there is no difference in a users understanding of video content at the media-level as a result of RoID technique used to present information. This finding complements previous work targeting non-RoI based video [GHI98; GHP00; GUL03] and suggests that variation of media-level video playback along RoI dimensions can take place without detrimental affecting the user's understanding of the video content.

An ANOVA test, with video clip as the independent variable and QoP-IA as the dependent variable was used to show whether video clip type significantly affected user understanding. It is noteworthy to observe that significant variation in user information assimilation {F(1,11) = 8.696 p<0.001} did occur as a result of the type of video clip being presented (see Figure 5.7b). In fact, results show that the type of video being presented is indeed more significant to the level of user information assimilation, i.e. understanding, than either the video quality variation, or RoID presentation technique.

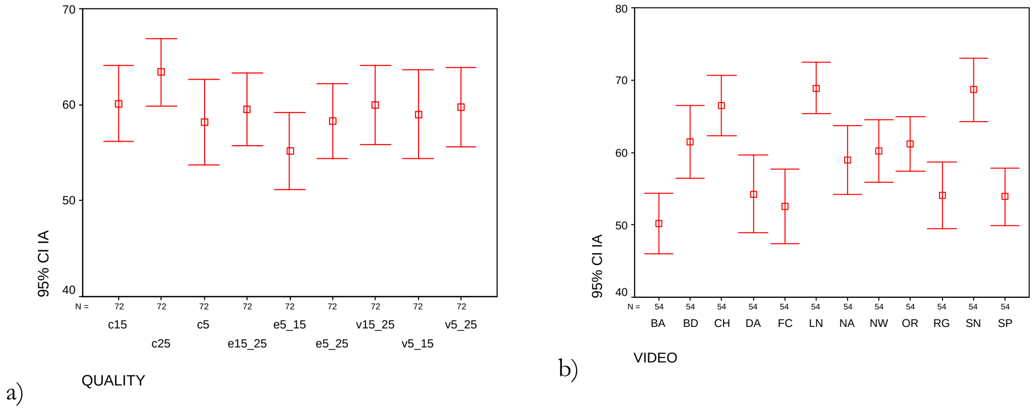


Figure 5.7: QoP-IA, dependent on quality (a) and video (b) type.

This finding is interesting, especially in the fields of advertising and education, as it implies that, to ensure users understand and assimilate the optimum level of information, the type of video being presented is significantly more important to the average media-level user information transfer than the QoS (see Figure 5.8).

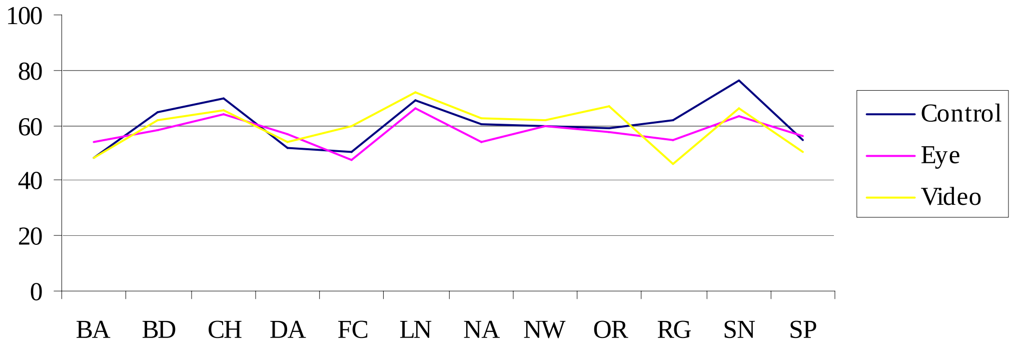


Figure 5.8: Average QoP-IA, independent of quality, shown for video clips.

5.4.2. Impact of RoID Quality on QoP-LoQ

User perceived level of quality (QoP-LoQ) defines the user's subjective opinion concerning the presentation quality of a particular piece of multimedia video, independent of the subject matter. A MANOVA test, with video variation quality and type of video as independent variables, and QoP-LoQ as a dependent variable, showed QoP-LoQ as being significantly affected by video variation quality {F(1,8) = 19.462 p<0.001}, the type of video {F(1,11) = 6.772 p<0.001}, as well as the combined effect of both factors {F(1,88) = 2.778 p<0.001} (see Figure 5.9b).

Results show that QoP-LoQ significantly impacts user perception within RoID presentation techniques: control videos, eye-based and content-dependent RoID videos (see 42a), as a result of video quality variation. This finding shows that the user is aware of the technical quality of video being presented. Indeed, MANOVA Post-Hoc Tukey-Tests showed a significant difference in QoP-LoQ between videos shown at 5 fps, compared to those shown at both 15fps {p=0.040} and 25 fps {p=0.031}. No significant difference was measured between QoP-LoQ when participants were shown videos at 15 and 25fps, suggesting that participants view 15fps and 25fps as being of similar perceptual quality (see Figure 5.9a). Indeed, users perceived level of quality for 25fps and 15fps control videos as being significantly different to all other quality definitions, with the exception of e25_15 and v25_15 (where only frame-rates of greater than 15fps were used). This supports the work of Wijesekera et al. [WIJ99], who shows that frame-rate should be maintained at or above 12 fps if the user QoP-LoQ is to be maintained.

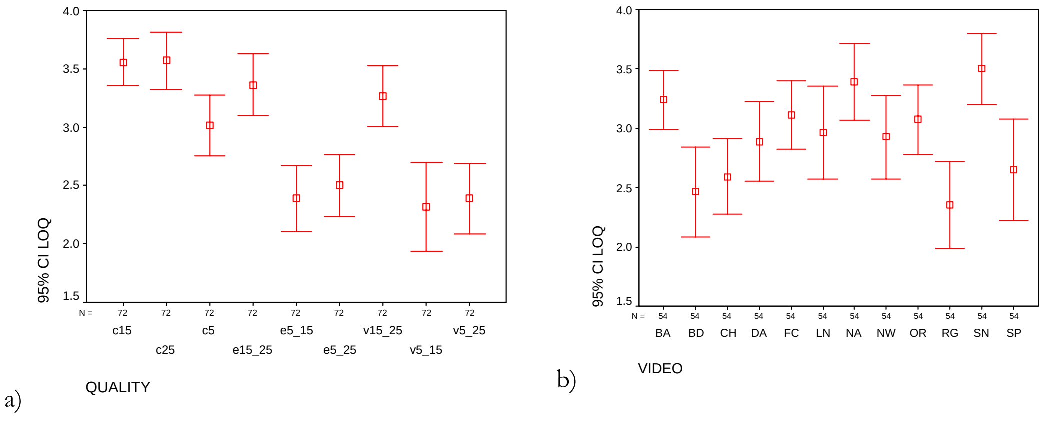


Figure 5.9: QoP-LoQ, dependent on quality (a) and video (b) type.

Post-Hoc Tukey-Tests showed significant differences in the level of QoP-LoQ between videos shown with a foreground/background combination of 25/15fps and all other RoID videos, independent of the display approach {e25/15 -

e25/5 : $p < 0.001$; e25/15 – e15/5 : $p < 0.001$; v25/15 - v25/5 : $p < 0.001$; v25/15 – v15/5 : $p < 0.001$}. This finding shows that using a multi frame-rate adapted RoIDs, with a background of less than 15fps, will negatively affect user QoP-LoQ, as shown in Figure 5.9a.

5.4.3 Impact of RoID Quality on QoP-LoE

QoP-LoE is the subjective level of enjoyment experienced by a user when watching a multimedia presentation. It is intuitive to assume that, as a result of personal preference, the type of video being presented to a participant will significantly affect a user's level of enjoyment. This is supported by our work {F(1,11) = 10.317 $p<0.001$} and can be clearly seen in Figure 5.10b. Interestingly, results showed that user level of enjoyment is not significantly affected by video quality variation {F(1,8) = 1.909 $p<0.056$} (see Figure 5.10a), even though we have already recognized that users are able to distinguish between the quality of video presentations. It appears that a conflict exist between video quality variation and video type, with user level of enjoyment being more significantly affected by the type of video being presented. This is interesting, especially in the field of entertainment and/or within bandwidth-constrained environments, as it suggests that changes in the presentation content at the media-level have more of an effect on QoP-LoE than a change to presentation technique (eye-based or content dependent RoIDs).

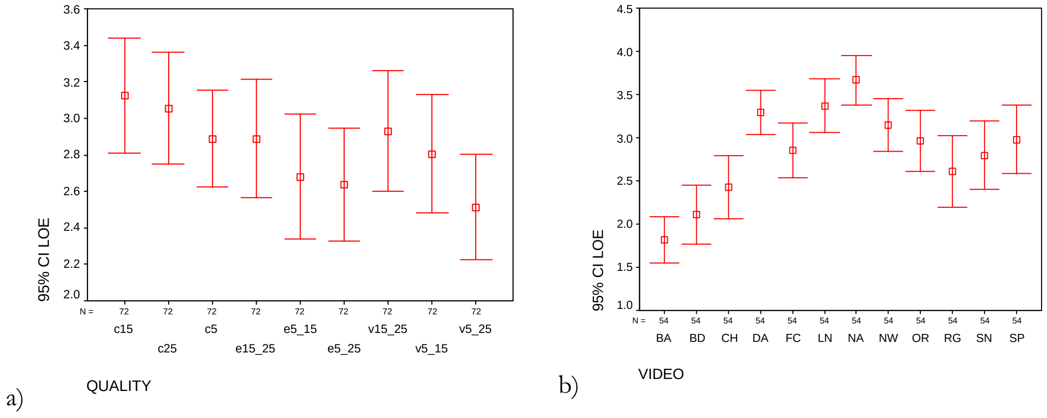


Figure 5.10: QoP-LoE, dependent on quality (a) and video (b) type.

5.5. Conclusions

In this chapter we implemented a RoI attentive display, which facilitates media-level variation as a result of both eye tracking- (user-perspective) and video content-based (technical-perspective) data, in order to measure the perceptual impact of media-level parameter variation; thus fulfilling research objective 2. Results showed that user QoP-IA is not affected by RoID presentation technique or video quality variation. Interestingly, QoP-IA is significantly affected by video clip, which is interesting, especially in the fields of advertising and education, as it implies that, to ensure users understand and assimilate the optimum level of information, the type of video being presented is significantly more important to the average media-level user information transfer than the QoS.

Results showed that the used RoID presentation technique and the type of video being presented both impact user QoP-LoQ. Our findings places considerable doubt upon whether frame rate based attentive displays can effectively maintain user QoP-LoQ, as they highlight that any use of frame rates less than 15 fps causes a significant reduction in user QoP-LoQ, i.e. that users are aware of video quality reduction.

Finally, results show that, although the type of video being presented significantly influences user QoP-LoE, both video quality variation and RoID presentation technique do not significantly affect user level of enjoyment. A conflict occurs between video quality variation and video clip type, with user level of enjoyment being more significantly influenced by the type of video. Findings suggest that the style of information transfer is more important to user enjoyment than the media-level QoS used to transfer data.

In conclusion, in this chapter we measure the impact of media-level quality parameter variation on user perception of multimedia quality by implementing an *attentive display*, which facilitates media-level variation as a result of both eye tracking- (User-Perspective) and video content-based (Technical-Perspective) data. We follow, in the next chapter, by measuring the impact of content-level quality parameter variation on user perception of multimedia quality, by manipulating display-type.

CHAPTER 6

Content-Level Quality Parameter Variation

In this chapter, we measure the impact of content-level quality parameter variation on user perception of multimedia quality, by manipulating display-type. To consider user-perspective content-level parameter variation, we intend to measure the impact of display-type on user perception of multimedia quality. Devices used include a fixed head-position eye-tracker, a traditional desktop limited mobility monitor, a head-mounted display, and a personal digital assistant. These devices represent considerable variation in screen-size, level of immersion, as well as level of mobility, which are all of particular importance in the fields of virtual reality and mobile communications. Technical-perspective content-level parameter variation is achieved through use of diverse experimental video material. As defined in our research aims, user perception of multimedia quality measurement will consider both user satisfaction (both with the QoS settings {media-level} and content enjoyment {content-level}), and user understanding. Although an eye-tracker will be used as a display device, no monitoring of user eye-gaze location data will be made at the content-level.

6.1. Introduction

The introduction of multimedia capabilities in mobile communications devices is a feature that, whilst increasing their allure, raises concerns as to whether user perception of multimedia quality varies as a result of the device type being used. Multimedia quality issues are of particularly important in mobile computing, since mobile devices are traditionally characterised by limited display capability [KIM01]. Indeed, the proliferation of mobile multimedia has brought a new dimension to the quality arena, since media must be suitably scaled in order to be appropriately displayed on the mobile display device. We contend therefore, that display device variation may impact user perception of multimedia quality, i.e. the ability to transfer information to the user, yet also provide the user with a certain level of satisfaction. We believe that if a device is perceived by the user as delivering low quality multimedia, users will rarely be convinced to pay for the privilege of using it, irrespective of its intrinsic appeal.

The structure of this chapter is as follows: In section 6.2, we provide a brief introduction to the output device types being considered (a fixed head-position eye-tracker, a traditional desktop limited mobility monitor, a head-mounted display, and a personal digital assistant), providing the reader with an understanding of the research that has been done, especially relating to the area of multimedia perception. In section 6.3, we describe the empirical study undertaken as part of our research, while Section 6.4 presents the main results obtained. Finally, in Section 6.5, conclusions are drawn.

6.2. Experimental Display Devices

In this section we aim to provide the reader with a concise introduction to the literature relating to the specific output display devices being considered in this chapter. This consequently allows us to consider the perceptual implication of varying level of mobility (see Table 6.1), since devices range from a fixed head-position eye-tracker, to a traditional desktop limited mobility monitor, to a head-mounted display (allowing greater autonomy of movement), through to a personal digital assistant (allowing full personal mobility).

Table 6.1: Varying Mobility of used Output Display Devices.

DEVICE	Eye Tracker	Generic Monitor	Eye Trek	Personal Digital Assistant
Mobility	Extremely limited mobility	Limited mobility due to the nature of desktop monitor	Provides mobility, yet gives restricted vision and requires supporting equipment.	Causes no mobility restriction. Can be used on the go.

6.2.1. Eye-Tracking

Considerable effort has been made, in Chapters 3, to introduce eye-tracking technologies. Subsequently, such work will not be repeated at this point. In summary: Eye-tracking systems can be used as a data gathering device or can provide the user with interactive functionality [ISO00; REI02]. Depending on the used equipment, eye-tracking devices can be considered as either intrusive or non-intrusive in nature [GOL02], can be developed as either pervasive [SOD02] or standalone systems and may have a level of immersion, which is perceived as being either high [HAY02] or low [PAS00].

An eye tracker system was chosen for our study, which limits user mobility, i.e. the user's head position must remain constant at all times, therefore providing a less autonomy of movement than traditional display devices.

6.2.2. Head Mounted Display

Head mounted displays have often been considered synonymous with virtual reality, however due to falling costs and improved technology, head-mounted displays devices are becoming more commercially available and have recently gained commercial importance for high street companies such as Olympus and Sony.

The head-mounted display is made of two canonical displays, and usually consists of two liquid crystal or cathode-ray tube display screens that are either mounted on a helmet or glasses-frame structure. There are several attributes that affect the usability of the head-mounted displays. Thus, head-mounted displays can be either binocular, showing the same image to both eyes, or stereoscopic in nature, showing different images to each eye. The choice between binocular or stereoscopic head-mounted displays depends on whether three-dimensional presentation of video media is required. Head-mounted displays use a range of display resolutions, however it is important to note that a trade-off exists between the display resolution used and the field of view, which in turn impacts the perceived level of experienced immersion [BOW02]. A low field of view decreases the experienced level of user immersion, yet a higher field of view involves spreading the available pixels, which can cause distortion on the picture. Finally, ergonomic and usability factors vary considerably between different devices. Issues such as display size, weight and adjustability of physical and visual settings all affect the usability of a particular head-mounted display for any specific task [BOW02].

Although there is now a wide range of head-mounted displays, there are several drawbacks that prevent their everyday popularity, including the current lack of available media that effectively facilitates immersive technology, e.g. full-motion, 3-D immersive video was developed within the last ten years [BOW02], yet is still commercially unavailable. The large and encumbering size is an important factor for the users of especially cathode-ray tube based displays [LAN97]. The subsequent physical limitations within the real world and reduced interactions with colleagues are also suggested reasons that prevent head-mounted displays from regular everyday popularity [LAN97]. Other concerns, such as hygiene and weight, also may have possible unknown long-term medical implications on the supporting muscles and even on the eyes.

Despite the usability drawbacks, head-mounted displays are used widely in a diverse series of research and development spheres, ranging from virtual reality environments to mobile wearable systems, which facilitate information access:

- There has been a great deal of effort in the virtual environment (VE) community towards developing new displays (e.g. [MEY99]) and improving existing display-types (e.g. [KIJ97]). However, there is limited work that objectively compares human behaviour and performance in different VE displays. Bowman et al. [BOW02] compared human behaviour and performance between a head-mounted display (HMD) and a four-sided spatially immersive display (SID). Bowman et al. showed that subjects have a significant preference for real-world rotation (rotation of the user in the real world) when using a HMD and virtual-rotation (rotation of the user's-perspective in the virtual world) when using the SID. Bowman et al. also found that females are more likely to choose real turns than males.

- Head mounted displays are a sub-set of wearable computer technology, which aims to allow hands free, mobile access to computer functionality. The motivation for hands free mobile computing devices is often varied, and ranges from individuals with restrictive physical disabilities [GIP96], to those working in dangerous or hazardous conditions [XYB03]. Moreover, the union of head mounted displays with wearable mobile devices, network technology, as well as effective input and output devices (touch pen, speech recognition inputs and interactive glove - as in the case of Xybernaut's Mobile Assistant [XYB03]), provides extremely adaptable mobile solutions, giving access to critical information, even in the most crowded public location [EBI02]. For example, the Smart Spaces project [PAB02] promises to implement anywhere / anytime automatic customisable, dynamically adaptable collaboration tools. In order to achieve these goals, the smart spaces project team have augmented virtual reality head mounted display systems with ubiquitous information access devices. The main driving force of this research project is to achieve information access anytime / anywhere, in order to support and improve the user's task performance.

A head-mounted display was chosen for our study, which although still limiting mobility (the user is restricted by limited vision and somewhat cumbersome equipment), facilitates a greater autonomy of movement than traditional display devices.

6.2.3. Personal Digital Assistant

Improvements in technology, especially in the wireless networking, have pushed the barriers of anywhere / anytime information access. Portable information access raises the need for portable information access devices, such as PDA (Personal Digital Assistants) and communicator devices (information-centric mobile phones that combines a fully featured PDA and mobile phone in one unit). Such devices promise to supplant the heavy desktop computer as ubiquitous technology, especially in educational and business environments [WEI98]. As mobile devices, PDAs inherit many of the problems associated with distributed and mobile computing systems [SAT01]. Distributed system problems include:

- *Remote tolerance*, such as protocol layering, the use of timeouts and remote procedure calls [BIR84].

- *Fault tolerance,* such as atomic transactions and distributed nested transactions [GRA93].

- *Remote Information access,* such as caching, distributed file systems and databases [SAT89].

- *Security,* related encryption-such as mutual authentication and privacy [NEE78].

Additional to the problems facing distributed systems, personal digital assistants also suffer from issues inherent to mobile computing devices [SAT01]. These include:

- *Mobile networking*, such as mobile IP [BHA96] as well as ad hoc protocols [ROY99].

- *Mobile information access,* such as disconnected operation [KIS92].

- *Support for adaptive applications,* such as transcoding by proxies [FOX96].

- *Location sensitivity,* such as location sensing and location-aware system behaviour [WAR97].

Personal digital assistants are also challenged by human-computer interaction (HCI) issues and ergonomic related concerns, such as small screen size, slow input facilities, low bandwidth, small storage capacity, limited battery lifetime and slow computer processor unit speed, which are all obstacles to the success of mobile and pervasive computing objectives and more importantly to user perception of multimedia quality [BUY00; FOX98; FUL01]. The increasing popularity and the above accumulation of problems have made PDAs a popular area of development. Studies / applications include:

- Jones et al. [JON99] studied the effect that screen size has on web-browsing related tasks. Their results show that users with small screens followed hyper-links less frequently than the users with a larger display unit, thus highlighting considerable changes in the method that users adopt when a user searches the internet for information using of small screen display (instead of a standard monitor).

- The Power Browser [BUY00], which was created to provide easy navigation in complex web sites using small screen mobile devices, such as a personal digital assistant. This application uses a hypertext transfer protocol proxy that receives the requests from the mobile user and, based on the request fetches of the user, dynamically generates a summary view to be transmitted back to the client. These summary web pages contain both link structure and contents of a set of web pages being accessed.

- Top Gun Wingman [FOC98], which is another transcoder that targets the Palm operating system. Although similar to the Power Browser, this application not only provides ease of navigation but instead converts the pages, images, and files (Zip / PalmDoc) to browser specific suitable formats.

- TV-Anytime [TVA03], which has applications that allow users to access their profiles remotely with a personal digital assistant and wireless Internet access [KAZ03]. Logged in users can search online databases for relevant television programmes, documentaries and movies and download to their home appliances. Using the application, the previews of the programmes can be watched online. Also, the user can set the length of the clip according to the network bandwidth and battery lifetime of the personal digital assistant.

Accordingly, PDAs represent, in our study, a truly mobile device, allowing the user full mobility of movement whilst viewing multimedia video presentations.

6.2.4. Perceptual Impact of Multimedia 'Quality of Service'

In this section we shall introduce the reader to perceptual studies that relate specifically to the display devices used in this study.

6.2.4.1. Eye-Tracking

Aspects concerning the perceptual impact of eye-tracking have already been extensively introduced in chapters 3 and 5, and will therefore not be elaborated upon in detail here. A summary of the area showed that:

- monitoring eye movements offers insights into user perception, as well as the associated attention mechanisms and cognitive processes, as the location of a users gaze corresponds to the symbol currently being processed in working memory [JUS76] and, moreover, that the eye naturally focuses on areas that are most likely to be informative [MAC70].

- objects in vision are determined from low-level visual primitives, such as edges, orientation, colour and motion [SAL66] (see Table 5.1). A user's point of gaze (i.e. the object currently being processed in working memory) is governed by cognitive high-level criteria, such as context, location, and the presence of people (see Table 5.2).

- four distinct looking states have been defined that summarise the cognitive state of the user [KAH73]: *Spontaneous looking*; *Task-relevant looking; Orientation of thought looking; Intentional manipulatory looking.*

6.2.4.2. Head Mounted Displays

A number of research studies exist looking at the symptoms related to head-mounted display usage, such as nausea [RAZ98], dizziness [COB95], headaches [KEN97] and eyestrain [KOL95]. However, to the best of our knowledge there has been no work done concerning QoP (i.e. the duality of both perception and satisfaction) of head-mounted display usage. Geelhoed et al. [GEE00], at Hewlett-Packard Laboratories in Bristol, investigated the comfort level of various tasks, such as text reading and video watching, whilst wearing different head-mounted displays. Geelhoed et al. identified that tasks that require more long-term attention, such as watching video, causes a greater level of discomfort to the user. It is important to note, however that Geelhoed et al. did not consider user information assimilation as part of its remit.

6.2.4.3. Personal Digital Assistant

To the best of our knowledge there has been no specific work done concerning the perceptual issues surrounding Personal Digital Assistants (PDAs), with the exception of Elting et al. [ELT02], which explores modality-combinations. Here Elting et al. [ELT02] looks at the effect that different output modality-combinations have on the devices' effectiveness to transport information and on the user's acceptance of the system being used. It uses three devices: a PDA, a television and a desktop computer to investigate whether the best modality-depends on the device. As test data, it uses a web based tourist guide that contains text and images. Their results showed that the most appealing form of information transfers combined picture, text and speech. However due to multi-modal cognitive load, especially when using a PDA, the most effective form of information transfer was shown to be just combined picture and speech.

6.3. Experimental Method

Our work implements content-level technical- and user-perspective quality parameter variation, by manipulating display-type and video content. Levels of informational transfer and user satisfaction are measured by incorporating the QoP concept, as defined in Chapter 3. In the following section we present specific information concerning QoP experiments and how the impact of content-level quality variation on user perception was measured.

6.3.1. Experimental Variables

Three experimental variables were manipulated in this chapter: type of device, multimedia video frame rate and multimedia content (video clip type). Accordingly, four types of display devices were considered in our experiments (representing varying levels of user mobility), and three multimedia video frame rates: 5, 15 and 25 frames per second. To ensure technical-perspective content-level quality parameter variation we used 12 video clips, which were identified by the user as being perceived significantly varied (see chapter 3). Eye tracking was no incorporated in this experiment, since eye-tracking prevents variation in display-type.

6.3.2. Experimental Participants and Set-up

The experiment involved 48 participants, who were aged between 18 and 56. Participants were evenly allocated into four different experimental groups (1, 2, 3 & 4) – 12 participants per group - to measure the impact that different display equipment has on user information assimilation and satisfaction. Within each respective group, users were presented the video clips using certain display equipment. Group 1 acted as a control group (standard mobility) and was therefore shown the video clips full screen using a normal 15 inch SVGA generic computer monitor enabled with a Matrox Rainbow Runner Video Card. Group 2 also viewed the video clips full screen using a computer monitor, however, the participants were simultaneously interacting with a Power Mac G3 (9.2) powered Arrington ViewPoint EyeTracker, used in combination with QuickClamp Hardware (see Figure 3.4), which provides limited mobility (see Table 3.3 for technical information concerning the ViewPoint EyeTrackter). Group 3 viewed the multimedia video clips using an Olympus Eye-Trek FMD 200 head-mounted display, which uses two liquid crystal displays and allows a greater autonomy of movement than a generic computer monitor. Each one of the displays contains 180,000 pixels and the viewing angle is 30.0° horizontal, 27.0° vertical. It supports PAL (Phase Alternating Line) format and has a display weight of 85g (Figure 6.1). Group 4 viewed the video clips using a Hewlett-Packard iPAQ 5450 personal digital assistant with 16-bit touch sensitive TFT liquid crystal display that supports 65,536 colours. The display pixel pitch of the device is 0.24 mm and its viewable image size is 2.26 inch wide and 3.02 inch tall. The PDA was using Microsoft Windows for Pocket personal computer 2002, operating system on an Intel 400 Mhz XSCALE processor and allows the user complete mobility. By default, it contains 64MB standard memory (RAM) and 48MB internal flash read-only memory (ROM). In order to complete this experiment a 128 MB secure digital memory card was used for multimedia video storage purposes (see Figure 6.2).

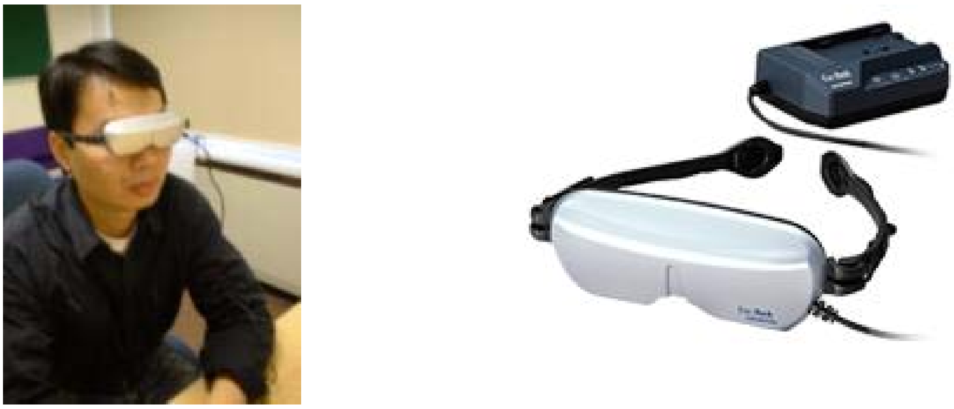


Figure 6.1: Head Mounted Display Device.

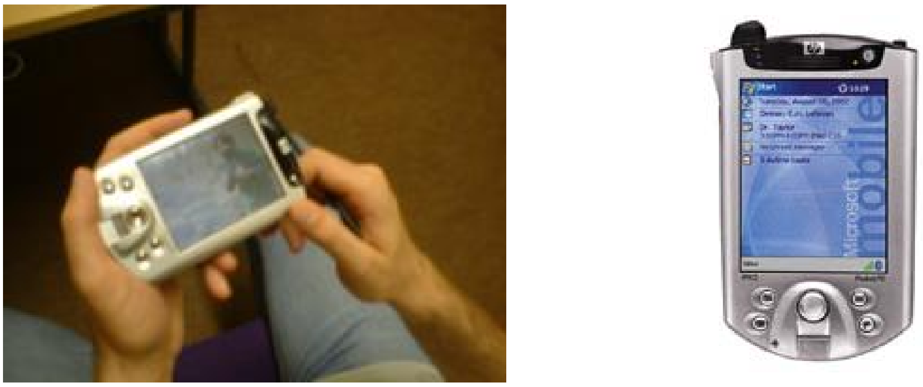


Figure 6.2: Personal Digital Assistant Device.

In addition, a pilot test study of 2 participants per display device (8 participants in total) was made to check and validate output experimental data. During this study, both of the test participants using the PDA commented that environmental noises interfered with the audio output. As we hoped to provide participants with a consistent audio level, headphones were used for all devices to limit interference from the surrounding environment.

In addition to participants viewing video material on different display devices – 12 participants per device, participants viewed video clips using one of three orders. Thus, each participant viewed four video clips at 5 frames per second, four video clips at 15 frames per second, and four video clips at 25 frames per second, with the order as defined in Table 6.2. Consideration of frame rate manipulation reduces the minimum size of sample to four participants. Although ideally a minimum sample size of six participants should be used, we were limited in our research by the practical number of people that were available. Using a reduced sample size in this way reduces the reliability of results concerning the impact of frame-rate on user perception of quality, yet it does not impact results relating to device variation or video clip type.

Table 6.2: Frame rate and video order presented to experimental groups.

Video	Content 1	Content 2	Content 3
BA	5	15	25
BD	25	5	15
CH	15	5	25
DA	25	15	5
FC	5	25	15
LN	5	15	25
NA	15	25	5
NW	5	25	15
OR	15	25	5
RG	25	5	15
SN	15	5	25
SP	25	15	5

6.3.3. Experimental Process

Independent of the display device being used, an experimental process was used that is consistent with that defined in section 3.6. The three video orders used in this experiment are defined by columns 'Content 1', 'Content 2' and 'Content 3' (see Table 6.2).

6.4. Results

6.4.1. Impact of Display-type on QoP-IA

Variation of device type was used in this study to facilitate content-level user-perspective quality parameter variation. Device variation also assists the identification of whether any significant changes occur to user QoP, due to changes in level of display mobility (see Table 5.2). To check the effect of device type on information assimilation (QoP-IA), we used a one-way AVOVA test, with QoP-IA as the dependent variable and device type as the independent variable. Results identified that display device variation causes significant variation in user QoP-IA {F(1,3) = 3.048, p=0.028}. ANOVA Post Hoc Tukey tests showed that a significant difference occurred between QoP-IA for participants using eye-tracker and head-mounted display devices {p= 0.018}. The head-mounted display and eye-tracker, were identified as respectively the best and worst devices for user video information assimilation.

We believe that the reason for the difference in user QoP-IA is due to the level of immersion available to the user whilst using the two devices. The Olympus Eye-Trek head-mounted display is designed to simulate a 52-inch display monitor, thus proving a high level of user visual immersion. Head-mounted displays also allow full head movement without changing the relative position of the screen and the eye. In comparison, the Arrington ViewPoint EyeTracker is used in combination with QuickClamp Hardware and headphone speakers, which intrusively restricts the movement of the user's head. Although restricted head movement is vital to this specific eye-tracker device to map and interpret eye-gaze location, it is intrusive and far from conducive to user immersion. Additional factors, such as a smaller perceived display screen (15 inch generic monitor), as well as the users' conscious awareness of the eye-tracker device are all possible factors that reduce participant visual immersion.

An ANOVA test, with video frame rate as the independent variable and user QoP-IA as the dependent variable shows that user understanding is not significantly affected by variation of video frame rate. This finding supports previous chapters, as well as the conclusions of Ghinea and Thomas [GHI98], which demonstrated that a significant loss of frames (that is, a reduction in the frame rate) does not proportionally reduce the user's understanding of the multimedia presentation. Current distributed and mobile computing multimedia systems often judge quality in terms of quality of service (QoS) provision. The results of our work, however, suggest that a significant change of objective QoS settings does not necessarily significantly affect the users' ability to assimilate information from multimedia video content. This allows us to justify the use of a lower frame rate, and therefore enabling a reduced bandwidth requirement, for multimedia video presentations, if and only if the user assimilation of information is the primary aim of the multimedia presentation.

It is common sense that the type of video clip should affect the source of user QoP-IA, as video clip determines the distribution of QoP-IA questions used in our experiment (see Table 3.2), e.g. the band clip has no textual content, so no textual feedback questions were used. Interestingly, when QoP-IA was analysed as a percentage measure of the number of questions being asked, considerable variation was still observed in user QoP-IA {F(1,11) = 10.769 p<0.001}. To analyse the impact of specific information sources on user IA, whilst considering possible null values, non-parametric Kruskal-Wallis K-independence tests were used to check whether clip type significantly impacted video, audio and textual information assimilation. Results showed that the type of clip, independent of frame rate and display-type, significantly affects information assimilated from video {χ^2(11, N = 576) = 287.833, p < .005}, audio { χ^2(11, N = 576) = 413.210, p < .005} and textual { χ^2(11, N=576) = 427.643, p < .005, supporting the findings of chapter 3 (see Figure 6.3).

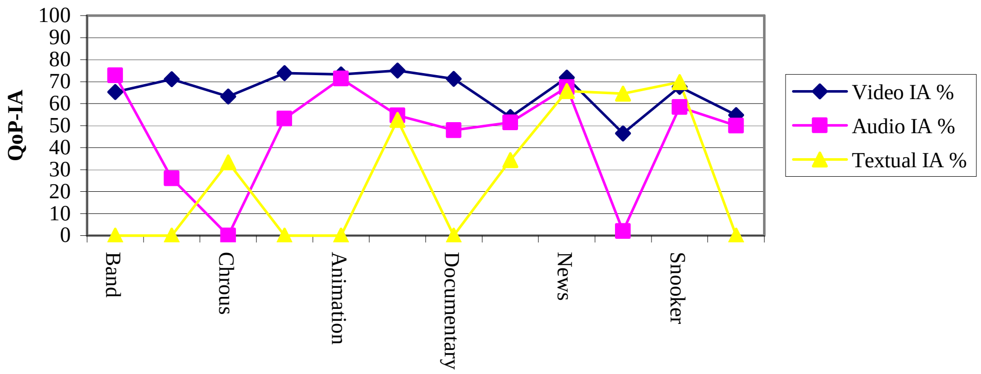


Figure 6.3: Average video, audio and textual information assimilation, independent of device type and frame-rate, for all video clips.

Our study shows that video clip type, used as part of a multimedia presentation, has more of a significant affect on a users' level of information transfer than either the frame rate, or display device type. The significant impact of the contents of the clip could be for a number of reasons: user preference leading to greater level of maintained attention, cultural level of pre-knowledge, clearer transfer of relevant information or even capable cognitive load. Although further work is required to determine the relationship between clip contents and level of information assimilation, we believe that this result supports our overall concern that when considering 'multimedia quality', we must consider two main facets: of service and of perception, since our work has implications on using purely objective testing when defining multimedia quality.

6.4.2. Impact of Display-type on QoP-LoQ

To measure the effect that device type has on QoP-LoQ, we used a one-way ANOVA, with device type as the independent variable and user QoP-LoQ as the dependent variables. A homogeneity of variance test showed that QoP-LoQ was not considered valid for AVONA analysis, since results were not evenly distributed around the mean. Consequently a Kruskal-Wallis K-independence non-parametric test was used to check the affect of device type on user QoP-IA. Results showed that device type significantly affected user QoP-IA {$\chi^2(3, N = 576) = 11.578, p= .009$}. Specifically, a significant difference was measured between mean QoP-LoQ for control (3.05) and head-mounted display (2.63) participant groups. It is important to note that head mounted displays, despite facilitating the greatest level of video information assimilation, are perceived by users as sustaining the worst QoP-LoQ (see Figure 6.4).

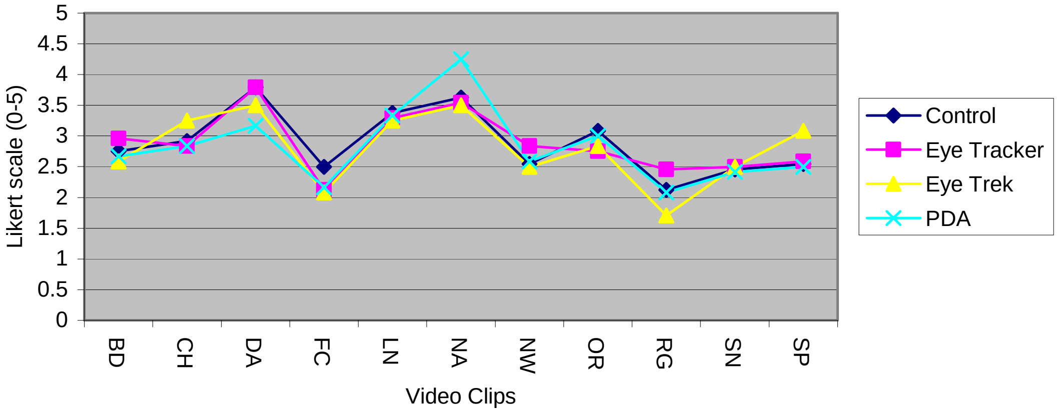


Figure 6.4: Average perceived QoP-LoQ, for all frame rates, across all video clips.

We believe that a significantly lower perceived level of quality may be due to one of two specific issues. The first proposed reason why head mounted displays negatively impact the user's QoP-LoQ is as a result of increased level of video immersion. As previously stated, a trade off exists between the resolution used and the field of view. A low field of view decreases the experienced level of user immersion, yet a higher field of view involves spreading the available pixels, which can cause distortion on the picture. The Olympus Eye-Trek head-mounted display is designed to simulate a 52-inch display monitor, providing the user with a high level of video immersion. As consistent video clips were used for all devices, we suggest that pixel distortion occurred as a result of a higher field of view, thus resulting low user QoP-LoQ. It is interesting to note that users viewing exactly the same videos on the 2.26 x 3.02 inch PDA screen perceived videos as being comparatively better quality (QoP-LoQ). The second proposed reason why head mounted displays causes a reduction in user perceived level of quality is due to physical discomfort. Our findings corroborate those of Geelhoed et al. [GEE00] who showed that, whilst using a head-mounted display, tasks requiring more long-term attention, such as watching a video, results in a greater level of discomfort to the user. Irrespective of the reason, the reduction of perceived quality has interesting implications on the future use of head mounted displays. We consider the users' satisfaction as being essential to any quality definition. Indeed, we are of the opinion that it is the person and not the machine or the underlying technology which is the ultimate determinant of quality: if an application is perceived to deliver low quality, users will rarely be convinced to pay for the privilege of using it, irrespective of its intrinsic appeal.

An ANOVA was used, with video frame rate as the independent variable and QoP-LoQ as the dependent variable, to measure the impact of frame rate variation on user QoP-LoQ. Results showed that user QoP-LoQ was significantly affected by variation in frame rate {F(1,2) = 4.766, p=0.009}. This finding confirms previous results, which show that the user is aware if degradation in the objective level of quality.

Previous findings have shown that clip-type has significant implications on user QoP-IA. To identify the impact of clip type on QoP-LoQ, a one-way ANOVA was used, with clip type as the independent variable and QoP-LoQ as the dependent variable. Results showed that clip type causes no significant variation in user QoP-LoQ at the content-level, unlike both network- and media-level quality abstractions. Results imply that technical parameter variation at the content-level does not cause a disparity in user QoP-LoQ, i.e. participants' identity that all videos are of the same objective quality. This result is believed to be as a consequence of network- and media-level video content variation (i.e. delay, jitter and attentive display RoI manipulation). This finding suggests that video content variation is more easily identified for certain video clips. Consequently, this disparity in QoP-LoQ, as a result of video clip type, reflects the ability of specific video to mask network- and media-level video errors. For example: the bath advert and snooker clip appears to effectively mask video variation; the band and rugby clip (both highly dynamic videos) do not effective hide network- and media-level video variation. Video content variation was not made at the content-level, therefore does not significant impact user QoP-LoQ.

6.4.3. Impact of Display-type on QoP-LoE

To determine the impact that device-type, frame rate and video clip has on QoP-LoE, we used a MANOVA, with device type, video frame rate and video clip type as the independent variables and user level of enjoyment (QoP-LoE) as the dependent variable. Results showed no significant difference in user QoP-LoE, as a result of device type or video frame rate. This finding implies that, at the content-level, a user's enjoyment is not affected by variation in display device or frame rate. Interestingly, results showed participant QoP-LoE to be significantly affected by video clip type {F(1,11) = 9.676, p<0.005} (see Figure 6.5).

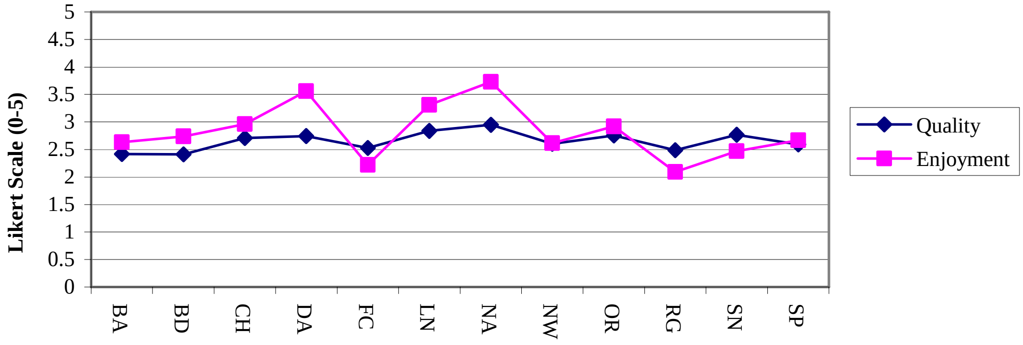


Figure 6.5: Average perceived level of Quality and Enjoyment, for all frame rates and devices, across all video clips.

As each specific participant has his / her own viewing preference, it therefore seems reasonable that the video clip type significantly affects a users' level of enjoyment. This result implies that a users QoP-LoE is clip dependent. Consequently this implies that, at the content-level, users are able to distinguish between their perception of 'quality' (QoP-LoQ) and their subjective appreciation of a video clip (QoP-LoE).

6.5. Conclusion

Our study measures the impact of content-level quality parameter variation on user perception of multimedia quality, by manipulating display-type. Quality parameter variation in this study was achieved, by showing participants a range of video clips at 3 different frame-rates (5, 15 and 25 frames per second), over a range of different display devices.

We propose that the perceptual effect of different device types cannot be generalised by obvious division into defined groups, such as mobile and non-mobile computing. The impact of device type should therefore be considered individually. Results show that the display device type used to watch a distributed multimedia video, significant impacts the user QoP-IA. Moreover, a significant difference was measured between the head-mounted display (HMD) device and eye-tracking device, which were respectively the best and worst devices for user information assimilation. We suggest that the reason for the difference in user QoP-IA is due to the level of immersion available to the users whilst using the two devices, with high-immersion devices (i.e. the HMD) facilitating a greater level of information assimilation. Although variation in device type does not significant impact user level of enjoyment, HMDs were found to significantly lower overall user perceived level of video quality, despite enabling the greatest level of video information transfer. If a device is perceived to deliver low quality, despite its ability to improve the transfer video information, we believe that users will rarely be convinced to pay for the privilege of using it. This conclusion has possible implications on the future of fully immersive head-mounted display devices and may result in a slow-down of commercial acceptance. We believe that this reduction in QoP-LoQ is due to pixel distortion as a result of a higher field of view and highlights the information/satisfaction compromise of using a head-mounted display system.

In addition, results show that: user QoP-IA is not significantly affected by a considerable loss of frames (that is, a reduction in the frame rate). This finding justifies the reduction in bandwidth allocation, if and only-if user QoP-IA is the primary aim of the multimedia presentation; video clip type has a more of a significant affect on the users' level of information assimilation than the frame rate and display device type; and finally that, although user QoP-LoQ is significantly affected by variations in frame rate and display device being used, it is not significantly affected by the video type. This implies that users are able to effectively distinguish between their subjective enjoyment of a video clip, and the video level of 'quality'.

In conclusion, in this chapter we measure the impact of content-level quality parameter variation on user perception of multimedia quality, by manipulating display-type and video clip type. In the final chapter of our study we summarise the research domain, state our research contributions and review our research findings.

CHAPTER 7

Conclusion

7.1. Research Domain

Distributed multimedia quality, in our perspective, is deemed as having two main facets: *of perception* and *of service*. The former quality facet (QoP - Quality of Perception) considers the user-perspective, measuring the infotainment aspect of the presentation, i.e. the ability to transfer information to the user, yet also provide a level of satisfaction. The latter quality facet (QoS – Quality of Service) characterises the technical-perspective and represents the performance properties that multimedia technology is able to provide. To effectively consider the contributions of previous studies, we extended Wikstrand's quality model [WIK03] in order to incorporate both user and technical quality perspectives. This showed that previous work, involving quality variation and measurement at the three levels of quality abstraction (network-, media- and content-levels) and the two quality perspectives (the technical- or user-perspectives) identified, failed to extensively measure the user's perception of multimedia quality. Moreover, although a number of studies consider the information duality of user perception [APT95; GHI98; GUL03; PRO99], no studies consider both how the multimedia presentation was assimilated / understood by the user at the content-level, but also examines the user's satisfaction (both his/her satisfaction with the objective QoS settings {media-level} and enjoyment concerning the video content {content-level}).

In light of these findings, our research defined the following research aim: to extensively consider the user's perception of distributed multimedia quality, by adapting relevant technical- and user-perspective parameters at all quality abstractions: network-level (technical-perspective), media-level (both technical- and user-perspectives), and content-level (both technical- and user-perspectives). In addition, we stated that the user-perspective must consistently consider both how the multimedia presentation was assimilated / understood by the user at the content-level, yet also the user's satisfaction (both his/her satisfaction with the objective QoS settings and enjoyment concerning the video content).

To achieve this research aim, a series of three main investigations were implemented, structured along the Network-, Media- and Content-levels of our model, each targeting a major research objective of our study.

- **Objective 1: Measurement of the perceptual impact of network-level parameter variation.** To consider network-level technical parameter variation we measured the impact of delay and jitter on user perception of multimedia quality. Although the impact of delay and jitter have been considered by other authors, previous studies fail to considering both level of user understanding (information assimilation) and user satisfaction (both of the video QoS and concerning the content of the video).

- **Objective 2: Measurement of the perceptual impact of media-level parameter variation.** Attentive displays monitor and/or predict user gaze, in order to manipulate allocation of bandwidth, such that quality is improved around the point of gaze [BAR96]. Attentive displays offer considerable potential for the reduction of network resources and facilitate media-level quality variation with respect to both video content-based (technical-perspective) and user-based (user-perspective) data. In order to measure media-level parameter variation, in respect of both technical- and user-perspectives, we measured the impact of a novel RoI attentive display system, which was developed to produce both video content- and user- dependent output video.

- **Objective 3: Measurement of the perceptual impact of content-level parameter variation.** To consider user-perspective content-level parameter variation, we measured the impact of various display-types on user perception of multimedia quality. Devices used include a fixed head-position eye-tracker, a traditional desktop limited mobility monitor, a head-mounted display, and a personal digital assistant. Technical-perspective content-level parameter variation was achieved through use of diverse experimental video material.

To ensure that perceptual measurement: i) was consistent, ii) considered the infotainment duality of distributed multimedia and iii) ensured that satisfaction considered both the user's satisfaction with the objective QoS settings, as well as the enjoyment concerning the video content, an appropriately adapted version of Quality of Perception (QoP) was used in all investigations. QoP uses user 'information assimilation' (QoP-IA) and user 'satisfaction' (QOP-S) to determine the perceived level of multimedia quality. QoP-S is subjective in nature and in our study consists of two component parts, targeting objective perceptual quality at the both media- and content-levels respectively: QoP–LoQ (the user's judgement concerning the objective Level of Quality assigned to the multimedia content being visualised) and QoP–LoE (the user's Level of Enjoyment whilst viewing multimedia content). To ensure content-level technical-perspective parameter variation, as well as the extensive usability of QoP-IA, it was important that the video material was perceived as being varied in nature. Accordingly, a short study was made that validated the use of video material initially chosen by Ghinea and Thomas [GHI98; GHP00], which was specifically chosen to cover a broad spectrum of infotainment. In addition to QoP, eye-tracking was incorporated in our work as experimental questionnaires do not allow continuous monitoring of user attention.

7.2. Research Contributions

The main contribution of our work stems from the defined research aims and objectives. Accordingly, in our work we have extensively considered the user's perception of multimedia quality, by adapting relevant technical- and user-perspective parameters at all quality abstractions: network-level (technical-perspective), media-level (both technical and user-perspectives), and content-level (both technical- and user-perspectives), ensuring that user perception of multimedia quality is considers both how the multimedia presentation was assimilated / understood by the user at the content-level, yet also examined the user's satisfaction (both his/her satisfaction with the objective QoS settings and level of enjoyment concerning the video content).

In support of our main contribution, the following sub-contributions were made:

- We extended Wikstrand's multimedia quality model [WIL03] in order to better reflect the infotainment duality of distributed multimedia. To the best of our knowledge, prior to our work, no multimedia quality models differentiated quality parameter variation and measurement in line with the infotainment duality of multimedia. With a growing research focus on understanding distributed multimedia quality, especially with reference to the user-perspective, it is important that standard distributed multimedia quality models are used to ensure consistency and comparison between research studies. Accordingly, we submit our extended model as an effective means of extensive quality paradigm definition and evaluation.

- We adapted Quality of Perception (QoP) [GHP00] to ensure that user satisfaction measurements considered both the user's satisfaction with the objective QoS settings and a user's level of enjoyment concerning the video content. Although QoP has been successfully used in previous studies, QoP has never been used to effectively measured user satisfaction at both media- and content-levels. Accordingly, by adapting QoP-S factors, we have facilitated a more extensive assessment of the user-perspective (see Figure 2.11).

- We incorporated eye-tracking in QoP experiments, in order to provide a better understanding of the role of the human element in the reception, analysis and synthesis of multimedia data. QoP questionnaires do not allow the continuous monitoring of user focus. Moreover, it can be argued that questionnaires do not conclusively highlight the points where information assimilation occurs, or whether the information was assimilated from the presentation at all. Consequently, use of eye-tracking provides a more conclusive solution to this dilemma. Prior to our work, eye tracking had not been used in support of QoP experiments.

- We developed unique median- and fixation map-based eye tracking data analysis methods, which when combined, facilitate analysis of 4-dimensional eye-tracking data: i.e. x-coordinate values, y-coordinate values, the number of samples showing the importance of an image region, and time. The majority of eye tracking research, with the exception of dedicated gaze-contingent display systems [PAR00], rely on static visual stimuli, e.g. a picture or a static web page. Consequently, previous eye-tracking data analysis techniques only consider three data dimensions: x-coordinate values, y-coordinate values and the distribution of samples showing the importance of an image region. To the best of our knowledge, no precedent or standard approach exists for analysing data that relate to non-static visual stimuli, i.e. a multimedia video clip. We therefore propose the combined use of median- and fixation map-based eye tracking data analysis in order to access 4-dimensional eye-tracking data. In our work software was developed, which: uniquely synchronised multiple participant eye-tracking data across multiple video frames (see Appendix E), produced relevant output data files (see Appendices C and E), produced region-of-interest scripts and fixation maps [WOO02] (See Appendices F and L respectively) and finally, analysed pixel distributions over time to aid analysis of user attention across video content (see Appendix F).

- We measured the impact of network-level technical-perspective parameter variation (delay and jitter) on user perception of multimedia quality (see chapter 4). Although impact of delay and jitter has been considered by other authors, these other studies fail to consider both level of user understanding (information assimilation) and user satisfaction (both of the video QoS and level of enjoyment concerning the content of the video). Accordingly, our perspective is considered as more extensive than previous studies, as the user-perspective measurement encompasses the infotainment duality of a multimedia presentation.

- We implemented, to best of our knowledge, the first frame rate-based RoID (Region-of-Interest Display) system. Although previous resolution-based RoID systems have been implemented [OSM98b], resolution gaze-contingent displays were shown to negative impact the users' ability to identify important information in peripheral areas of vision [LOS00; REI00]. Consequently, as frame rate is a major factor influencing multimedia bandwidth requirements, and is of primordial importance because of to its scarcity in distributed multimedia environments, a frame rate-based RoID system was used to measure the impact of media-level variation on user perception of multimedia quality, i.e. video content- (technical-perspective), and user-dependent (user-perspective) frame rate-based RoID video (see chapter 5). As our investigation manipulates both technical- and user-perspective media-level quality parameters, as well as measuring the user-perspective at both media- and content-levels (encompassing the infotainment duality of multimedia), the results of our work more extensively consider the user-perspective at the media-level than previous studies.

- To effectively deliver region of interest (RoI) data to the frame rate-based RoID software (See Appendix M), we developed RoI scripts (See appendix L). RoI scripts provide a flexible method of defining user RoI data, which in our study ensured format consistency between video content- (technical-perspective), and user-dependent (user-perspective) RoI data. To the best of our knowledge, no alternative method has been proposed for storing RoI information. Standardisation of input, for all RoI-based attentive display systems, would allow consistent comparison between multiple display techniques. Consequently, we propose our RoI scripts as an effective means of representing RoI data for all RoI-based attentive display systems.

- We measured the impact of content-level technical- and user-perspective parameter variation (video type and device type respectively) on user perception of multimedia quality (see chapter 6). Devices used in our work included a fixed head-position eye-tracker, a traditional desktop limited mobility monitor, a head-mounted display, and a personal digital assistant. As well as manipulating both technical- and user-perspective content-level quality parameters, our study measuring the user-perspective at both media- and content-levels (encompassing the infotainment duality of multimedia). To the best of our knowledge no other studies has extensively considered the impact of content-level parameter variation on the user's perception of multimedia quality.

7.3. Research Findings

Our research findings can be effectively divided into two sections. These sections correspond to the two assessment techniques used, in our work, to measure the user's perception of distributed multimedia quality - the user-perspective, i.e. Quality of Perception (QoP) and Eye tracking.

7.3.1. Quality of Perception

QoP was used in our study to extensively consider the user's perception of multimedia quality (the user-perspective). Accordingly, three main investigations were used to measure QoP-IA (the user's ability to assimilate information) and user QoP-S (the user's satisfaction), as a results of relevant technical- and user-perspective parameter variation, made at the network-level (technical-perspective), the media-level (both technical- and user-perspectives), and the content-level (both technical- and user-perspectives) respectively. QoP-S is subjective in nature and in our study consists of two component parts: QoP–LoQ (the user's judgement concerning the objective Level of Quality assigned to the multimedia content being visualised) and QoP–LoE (the user's Level of Enjoyment whilst viewing a multimedia content). As defined in our study objectives:

- To consider network-level technical-parameter variation, we measured the impact of delay and jitter on user perception of multimedia quality (See Chapter 4).
- To measure media-level parameter variation, in respect of both technical- and user-perspectives, we measured the impact of a novel attention-based display system, which was developed to produce both video content- and user-dependent output video (See Chapter 5).
- To consider user-perspective content-level parameter variation, we measured the impact of varied display-type on user perception of multimedia quality (See Chapter 6). Technical-perspective content-level parameter variation was achieved through use of diverse experimental video material.

In addition to abstraction-level quality parameter variation, we also measured impact of video frame rate and video clip type at each level of our quality model.

Table 7.1: A summary of our QoP finding. (✗ - no significant difference; ✓ - signficant difference).

		QoP-IA	QoP-LoQ	QoP-LoE
Network Level	Delay	✗	✓	✓
	Jitter	✗	✓	✓
	Frame Rate	✗	✓	✓
	Video Clip	✓	✓	✓
Media Level	Attentive Display	✗	✓	✗
	Frame Rate			
	Video Clip	✓	✓	✓
Content Level	Device Type	✓	✓	✗
	Frame Rate	✗	✓	✗
	Video Clip	✓	✗	✓

The findings of our work (See Table 7.1) highlight a number of important issues relating to the effective provision of user-centric quality multimedia. These issues will now be discussed.

- A significant loss of frames (that is, a reduction in frame rate) does not proportionally reduce the user's understanding of the presentation (see Table 7.1). This finding supports the conclusions of Ghinea and Thomas [GHI98] and justifies the reduction in bandwidth allocation, if and only-if user QoP-IA (information assimilation / understanding) is the primary aim of the multimedia presentation.

- A significant loss of frames, specifically below 15fps, was found to significant impact user QoP-LoQ. This finding supports the work of Wijesekera et al. [WIJ99], who showed that frame-rate should be maintained at or above 12 fps if the user perception of multimedia quality is to be maintained. Interestingly, this finding also raises considerable concerns regarding the usability of frame rate based attention display systems, since our findings show no positive benefits to frame rate based attentive displays.

- Video clip type significantly impacts user QoP-IA (Information Assimilation). Variation in user QoP-IA shows that the level of information assimilated significantly varies across the range of experimental video material. As the informational content of video determines the use of QoP-IA questions, and ultimately the reliability of QoP-IA, this finding supports the use of QoP-IA at each of the quality-abstractions.

- Video clip type significantly impacts user QoP-LoE (Level of Enjoyment). Variation in user QoP-LoE, shows that certain videos (FC, LN, DA in our study) were perceived as being more enjoyable. This finding is of interest, especially in the fields of advertising and education, as it implies that, the type of video is more significantly important to the users' level of enjoyment than implementing certain quality parameter variation, e.g. variation in the device type. Further work is required to fully understand the relationship between video content and user enjoyment, yet this aim lies outside the scope of our study.

- User QoP-IA is significantly affected by variation in content-level parameter variation (device type), yet is not significantly affected by network-level and media quality parameter variation. Results show that the display device, used to watch a distributed multimedia video, significant impacts user QoP-IA. A significant difference was measured between the head-mounted display (HMD) device and eye-tracking device, which were identified as respectively the best and worst devices for user information assimilation. We believe that the reason for the difference in user QoP-IA is due to the level of immersion, with high-immersion devices (i.e. the HMD) facilitating a greater level of information assimilation. Although variation in device type does not significant impact user level of enjoyment, HMDs were found to significantly lower overall user perceived level of video quality (QoP-LoQ), despite enabling the greatest level of video information transfer. We suggest that this reduction in QoP-LoQ is due to pixel distortion as a result of a higher field of view and highlights the information/satisfaction compromise of display systems, i.e. a higher field of view provides a higher QoP-IA, yet provides a lower QoP-LoQ (and visa-versa). This conclusion has possible implications on the future of fully immersive head-mounted display devices, as we believe that any device that is perceived to deliver low quality, despite its ability to improve the transfer video information, will rarely be commercially accepted by the user.

- User QoP-LoQ is significantly affected by network-, media-, and content-level quality parameter variation, i.e.: delay, jitter, attentive display RoI manipulation, and device type. This finding shows that participants can effectively distinguish between a video presentation with and without error. This finding supports [WIJ96], who showed that the presence of even low amounts of error results in a severe degradation in perceptual quality. Consequently, it is essential to identify the purpose of the multimedia when defining appropriate QoS provision, e.g. applications relying on user perception of multimedia quality should be given priority over and above purely educational applications.

- User QoP-LoQ is significantly affected at the network- and media-level, yet QoP-LoQ is not significantly affected by video clip type at the content-level. This result is believed to be as a consequence of network- and media-level video content variation (i.e. delay, jitter and attentive display RoI manipulation). This finding suggests that video content variation is more easily identified by users in certain video clips. Consequently, this disparity in QoP-LoQ, as a result of video clip type, reflects the ability of specific video to mask network- and media-level video variation errors. For example: the bath advert and snooker clip appears to effectively mask video variation errors mask; the band and rugby clip (both highly dynamic videos) do not effective hide network- and media-level video variation errors. Video variation was not made at the content-level, therefore does not significant impact user QoP-LoQ. This finding also supports previous finding that participants can effectively distinguish between a video presentation with and without error.

- User QoP-LoE is significantly affected by network-level quality parameter variation (jitter and delay), yet is not significantly affected by media-level and content-level quality parameter variation (attentive display RoI manipulation and display-type). This findings support Procter et al. [PRO99], who observed that degradation of network-level QoS has a greater influence on a subjects' uptake of emotive / affective content than on their uptake of factual content. This result has serious implications on the effective provision of user-centric quality multimedia, implying that: If you wish to ensure user QoP-IA, then network-level quality parameter variation should be used; If you wish to maintain user QoP-LoE, then content-level quality parameter variation should be used.

7.3.2. Eye Tracking

Eye-tracking was used in our study to help identify how gaze disparity in eye-location is affected at the network-level. Eye-tracking was measured at the network-level, and due to the complexity of eye-tracking data, was analysed separately to QoP data. In our study we have introduced two eye tracking data analysis methods (median and fixation-map eye tracking analysis), which when combined consider the 4-dimensions of eye-tracking data: i.e. x-coordinate values, y-coordinate values, the distribution of samples showing the importance of an image region, and time. Both methods, although not ideal, provide important information concerning user attention and video eye-path.

7.3.2.1. Median Eye Tracking Analysis

Median eye tracking analysis was used in our study to provide statistical analysis of eye-tracking data at the network-level. Mapped median eye tracking analysis considers two of the four possible data dimensions (a single coordinate value and time), yet reduces analysis complexity and facilitates statistical analysis.

Results showed that frame rate alone does not impact the user video eye-path, i.e. similar trends in median eye movement occur for groups of people when shown the same video content at different frame. The use of additional quality parameter variation (e.g. jitter / delay) results in disparity in user video eye-paths. Interestingly this only happens for certain video (BD, CH, DA, FC), which will be considered further in the following section.

7.3.2.2. Fixation Map Eye Tracking Analysis

Fixation Maps were used over time to better understand how quality parameter variation (i.e. specifically jitter and delay) causes disparity in video user eye-path. Fixation map eye tracking analysis considers two eye tracking data dimensions (the distribution of samples / the importance of an image area, and time).

Results showed that variation in video eye path occurs when: i) no single / obvious point of focus exists; or ii) when the point of attention changes dramatically. It is important to note that dynamic video content does not negatively impact user region of interest, as long as a small single point of focus exists, e.g. in the rugby clip, it is only after the score (i.e. after the ball has been removed), that the highest difference in video eye-path exists. If a full screen or large single stimuli exists then disparity in user eye-path occurs, e.g. the man wiping the bath with the sponge in the bath advert video clip.

Median eye tracking analysis showed an increased disparity in user video eye-paths, as a result of video quality parameter variation. Interestingly, disparity only existed for BD (Band), CH (Chorus), DA (Animation) and FC (Weather forecast) videos, which, for the majority of the video, do not have a single / obvious point of focus. Instead multiple conflicting points of focus exist in BD, CH, DA and FC, which ultimately causes variation in user eye-path, as a result of delay and jitter. Moreover, the point of focus in BD and DA video changes dramatically as a result of fast scene changes or multi person dialogue, i.e. video material does not ensure smooth pursuit eye movement, thus resulting in problems identifying and tracking important points-of-focus.

7.4. Potential Research Implications and Applications

Although research contributions and findings have already been stated, we have not explicitly stated the potential implications and applications of our work on future multimedia research.

7.4.1. Unified Assessment Perspective

Traditionally numerous 'quality' parameters were used to measure the quality of distributed multimedia. Results from this research have specifically identified measurable variation in optimum user multimedia perception as a result of the participant age, individual viewing preference, as well as the type of physical equipment and level of mobility being implemented during a multimedia presentation. Importantly we have introduced variation of parameters within the context of a proven model of multimedia quality, therefore providing a unified assessment perspective allowing multi-study comparison. Further assessment of specific quality parameters is vital to the more comprehensive understanding of user perception, however future assessment should be made in context of the quality model provided in our work (variation is at either the network-, media-, or content-level, from either a user- or technical perspective) so that end-to-end system perception can be achieved.

7.4.2. Personalisation of Media

Traditionally the perception of multimedia quality has been driven from a purely technical perspective, which although measurable dismisses the user. We believe that the specific user will not continue to support multimedia systems if they are perceived to be of bad quality. Moreover, by dismissing the user we ultimately risk ignoring accessibility concerns, by excluding access for users with abnormal perceptual requirements, e.g. the deaf. To comprehensively assess and incorporate specific user perception in multimedia technologies, addition research is required to understand the impact that multimedia setting and computer device variation has on user perception of multimedia quality. Such research facilitates the appropriate allocation of technical provision in context of the perceptual, hardware, physical and network requirements of a specific user - thus maximising the specific user's experience of quality. Personalisation of media streaming provides truly user-defined accessible multimedia, allowing the user to interact directly with systems on their own perceptual terms. Implied future research includes: a Personalised Perceptual Profile (P^3) to allow the user perspective to be defined; proper assessment methods for defining multimedia content to allow better perceptual based media adaptation; incorporation of Artificial Intelligence to facilitate the real-time adaptation of media streams for the personalised perceptual benefit of various users.

7.4.3. Eye-tracking Data Representation

Eye-tracking is 4-dimensional in nature (X-coordinate, Y-coordinate, number of samples and time), making statistical analysis and visual representation of eye-tracking data a recognised concern. As part of out work, we developed a number of rudimentary approaches for the interpretation and manipulation of eye-tracking data. As well as providing initial data representation techniques, which may be used where multi-dimensional data representation is a problem (e.g. informatics), finding support future work including: the implementation of a multi-dimensional environments (i.e. virtual reality) in order to facilitate the visualisation and manipulation of eye-tracking data; automated region of interest analysis, to support personalised perception definition or media adaptation.

7.4.4. Attentive Displays

Attentive displays use either information concerning the user's point of gaze (gaze-contingent), or information relating to the physical characteristics of the specific video (the Regions-of-Interests), to manipulate the allocation of display bandwidth, such that a greater level of 'quality' is provided within the centre of vision. In our work we developed a multi frame-rate region-of-interest display, which manipulated MPEG-1 video using both eye-tracking data and information concerning the video-content. Although in our work attentive displays were shown to have no perceptual benefit, a number of possible research areas have been identified including: using content defined MPEG-4 video screen descriptors instead of fixed shape region-of-interest manipulation in MPEG-1; a variable region-of-interest frame-rate ratio (5_15, 15_25, etc) that is based on the video content (specifically dynamic frame variation); as well as combined live and client animated video regions. Attentive displays have considerable potential in online and interactive video and gaming, especially in limited or reduced bandwidth environments, such as mobile devices.

7.5. Conclusion

Our work has shown that user perception of distributed multimedia quality (QoP) cannot be achieved by means of purely technical-perspective QoS parameter adaptation. Accordingly, the future of multimedia research contains both promise and danger for user-perspective concerns.

As previously stated, we believe that a user will not continue paying for a multimedia system or device that they perceive to be of low quality, irrespective of its intrinsic appeal. Consequently, if commercial multimedia development continues to ignore the user-perspective in preference of other factors, e.g. user fascination (i.e. the latest gimmick), then companies risk ultimately alienating the customer. Moreover, by ignoring the user-perspective, future multimedia systems risk ignoring accessibility issues, by excluding access for users with abnormal perceptual requirements, e.g. the deaf [GUL03].

If, by contrast, commercial multimedia development effectively considered Quality of Service parameter variation in context of a specific user's perception of quality, then multimedia provision would naturally aspire to facilitate optimum multimedia independent of the perceptual, hardware and network criteria, thus maximising user perception of quality. Furthermore, the development of user-specific personalisation and adaptation of multimedia streams offers the customer with truly user-defined and accessible multimedia, facilitating direct interaction with multimedia information on their own specific perceptual terms.

By providing an extensive study of the distributed multimedia quality, our work shows that the user-perspective is as critically important to distributed multimedia quality definition, as QoS technical parameter variation. In conclusion, although multimedia applications are produced for the education and / or enjoyment of human viewers, effective integration and consideration of the user-perspective in multimedia systems still has a long way to go…

BIBLIOGRAPHY

AHA93 Ahumada, A. J., (1993). Computational Image quality metrics: A review. *SID Symposium Digest, Vol. 24,* pp. 305-308.

AHN93 Ahumada, A. J., and Null Jr, C. H., (1993). Image quality: A multidimensional problem. *Digital Images and Human Vision*, Watson, A.B, (Ed.), MIT Press, pp. 141-148.

ALD95 Aldridge, R., Davidoff, J., Ghanbari, M., Hands, D., and Pearson, D. (1995). Recency effect in the subjective assessment of digitally coded television pictures. In *Proc. IPA,* Edinburgh, UK, pp. 336-339.

APT95 Apteker, R.T., Fisher, J.A., Kisimov, V.S., and Neishlos H., (1995). Video Acceptability and Frame Rate. *IEEE Multimedia*, *Vol. 2*, No. 3, Fall, pp. 32 - 40.

ARD94 Ardito, M., Barbero, M., Stroppiana, M., and Visca, M., (1994). Compression and Quality. *Proceedings of the International Workshop on HDTV '94,* Chiariglione, L., (Ed.), Torino, Italy, October 26 - 28, 1994, Springer Verlag, pp. B-8-2.

AYA98 Ayabe–kanamura, S. Schicker, I., Laska, M., Hudson, R., Distel, H., Koboyakawa, T., and Saito S., (1998). A Japanese-German cross-cultural study, *Chemical Senses, Vol. 23*, pp. 31-38.

AXE95 Axel, R., (1995). The molecular logic of Smell. *Scientific American, Vol. 10*, pp. 130-137.

BAL81 Baldwin, D., (1981) Area of interest: Instantaneous field of view vision model. *Image Generation / Display Conference II,* pp. 481-496.

BAL94 Baluja, S., and Pomerleau, D., (1994). Non-intrusive gaze tracking using artificial neural networks, *Research Paper CMU-CS-94-102,* School of Computer Science, Carnegie Mellon University, Pittsburgh PA, USA.

BAR96 Barnett, B. S., (1996) Motion Compensated visual pattern image sequence coding for full motion multi-session video conferencing on multimedia workstation. *Journal of Electronic Imaging, Vol. 5*, pp. 129-143.

BAU03 Baudisch, P., DeCarlo, D., Duchowski A. T., and Geisler, W. S., (2003) Focusing on the Essential: Considering Attention in Display Design, *Communications of the ACM*, *Vol. 46* (3), pp. 60-66.

BHA96 Bhagwat, P., Perkins, C., and Tripathi, S., (1996) Network Layer Mobility: An Architecture and Survey. *IEEE Personal Communications Vol. 3* (3), pp. 54-64.

BIR84 Birrell, A.D., Nelson, B.J., (1984). Implementing Remote Procedure Calls. *ACM Transactions on Computer Systems, Vol. 2*, (1), pp. 39-59.

BLA92 Blakowski, G., Hübel, J., Langrehr, U., and Mühlhäuser, M., (1992). Tools support for the synchronisation and presentation of distributed multimedia. *Computer Communications, Vol. 15,* No 10, pp. 611-618.

BLA96 Blakowski, G., and Steinmetz R., (1996). A Media Synchronisation Survey: Reference Model, Specification, and Case Studies. *IEEE Journal on Selected Areas in Communications, Vol. 14,* No. 1, pp. 5–35.

BLO03 Blochl, M., Rumetshofer, H., and Wob, W., (2003). Individualized e-learning systems enabled by a semantically determined adaptation of learning fragments *Proc. 14th International Workshop on Database and Expert Systems Applications*, pp 640–645.

BOH97 Bohannon, W., (1997). Buyers Guide to LCD and DLP Projectors. *Presentations Magazine,* pp. 53-64.

BOK00 Bouch, A., Kuchinsky, A., and Bhatti, N., (2000). Quality is in the eye of the beholder. *Proc. of the CHI 2000 Conference on Human Factors in Computing Systems*, The Hague, The Netherlands, pp. 297-304.

BOS00 Bouch, A., Sasse, M.A., and DeMeer, H., (2000). Of Packets and People: A User-Centred Approach to Quality of Service. *Proc. of IWQoS 2000*, Pittsburgh, PA, pp. 189-197.

BOU98 Bouch, A., Watson, A., and Sasse, M.A., (1998). QUASS - A Tool for Measuring the Subjective Quality of Real-Time Multimedia Audio and Video. Poster presented at *HCI '98*, Sheffield, England. Available.

BOU01 Bouch, A., Wilson, G. and Sasse M. A., (2001) A 3-Dimensional Approach to Assessing End-User Quality of Service. *Proc. of the London Communications Symposium,* pp.47-50.

BOW02 Bowman, D., Datey, A., Ryu, Y., Farooq, U., and Vasnaik, O., (2002). Empirical Comparison of Human Behavior and Performance with Different Display Devices for Virtual Environments. *Proc. of the Human Factors and Ergonomics Society Annual Meeting,* pp. 2134-2138.

BRU96a Bruce, V., Green, P., and Georgeson, M., (1996), "Visual Perception: Physiology, Psychology, and Ecology, Psychology Press.

BRU96b ________________ pp. 31.

BRU96c ________________ pp. 32.

BRU96d ________________ pp. 34.

BRU96e ________________ pp. 420.

BUY00 Buyukkokten, O., Garcia-Molina, H., Paepcke, A., and Winograd, T., (2000). Power Brower: Efficient Web Browsing for PDAs. In *ACM CHI 2000*, The Hague, Amsterdam, pp. 430-437.

BYR99 Byrne, M. D., Anderson, J. R., Douglass, S., and Matessa, M., (1999) Eye Tracking the Visual Search of Click-Down Menus. *Proc. Of ACM CHI '99,* Pittsburgh, Pennsylvania, USA, pp. 402-409.

CAL93 Caldwell, G., and Gosney, C., (1993) Enhanced tactile feedback (tele-taction) using a multi-functional sensory system. *Proc. IEEE International Conference on Robotics and Automation*, Atlanta, GA, pp. 955-960.

CAR98 Carney, T., (1998). Mindseye: A visual programming and modelling environment for imaging science. *Proc. SPIE, Vol. 3299*, San Jose, CA, pp. 48-58.

CAT92 Cater, J. P., (1992). The nose have it! *Presence, Vol. 1,* No. 4, pp. 493–494.

CAT94 Cater J. P., (1994). Smell/Taste: odours. *Virtual Reality,* pp 1781.

CCI74 CCIR (1974), Method for the Subjective Assessment of the Quality of Television Pictures, *13th Plenary Assembly, Recommendation 500, Vol. 11*, pp. 65–68.

CHI04 Chiariglione, L., (2004). The MPEG Home Page. Retrieved from the Moving Picture Experts Group (July 2004) http://www.chiariglione.org/mpeg/

CLA98 Claypool, M., and Riedl, J., (1998). End-to-end quality in multimedia application. In *Chapter 40 in Handbook on Multimedia Computing,* CRC Press, Boca Raton, FL.

CLA99 Claypool, M., and Tanner, J., (1999). The Effects of Jitter on the Perceptual Quality of Video, *ACM Multimedia'99 (Part 2),* Orlando, FL, pp. 115-118.

CLE92 Cleveland, D. and Cleveland, N., (1992). Eyegaze eyetracking system. *Imagina: Images Beyond Imagination. Eleventh Monte-Carlo International Forum* *on New Images*, LC Technologies, Inc., 4415 Glenn Rose Street, Fairfax, Virginia 22032, U.S.A.

COB95 Cobb, S., Nichols, S., and Wilson, J.R., (1995). Health and safety implications of virtual reality: *Proc. of FIVEi95 (Framework for Immersive Virtual Environments)*, University of London, pp. 227-242.

COH90 Cole B. L. and Hughes P. K., (1990). Drivers don't search: they notice. *Visual Search*, Brogan D., (Ed.), pp. 407-417.

COH92 Cohn, M.B., Lam, M., and Fearing, R.S., (1992) Tactile feedback for teleoperation. *Proc. SPIE Telemanipulator Technology*, Vol. 1833, Boston, Das, H., (Ed.), pp. 240-254.

COL90 Cole, G. R., Stromeyer, C. F. (3rd), and Kronauer, R. E., (1990). Visual interactions with luminance and chromatic stimuli. *Journal of the Optical Society of America A*, *Vol. 7* (1), pp. 128-140.

COM90 Comes, S., and Macq, B., (1990). Human Visual Quality Criterion. *SPIE Visual Communications and Image Processing, Vol. 1360*, pp. 2-7.

COO94 Coolican, H. (1994). Research methods and statistics in psychology. (4th edition). London: Hodder and Stoughton.

CZE02 Czerwinski, M., Tan, D.S., and Robertson, G.G., (2002). Women take a wider view. *Proc. of the CHI '02, MN, ACM NY,* pp. 195-202.

DAL93 Daly, S., (1993). The visible differences predictor; An algorithm for the assessment of Image fidelity. *Digital Images and Human Vision*, Watson. A. B. (Ed.), MIT Press, pp. 179-206.

DAV01 Davide, F., Holmberg, M., and Lundstrom, I., (2001). Virtual olfactory interfaces: electronic noses and olfactory displays. *Communications though virtual Technology: Identity, Community and Technology I the Internet Age*, Chapter 12, IOS Press, Amsterdam, 2001, pp. 193-219.

DEG66 De Groot, A. D., (1966). Perception and memory versus thought: Some old ideas and recent findings. *Problem solving: Research, method, and theory,* Klinmuntz, B. (Ed.), New York: John Wiley.

DEM75 de Monasterio, F. M and Gouras, P., (1975). Functional propertied of ganglion cells of the rhesus monkey retina. *Journal of Physiology, Vol. 251,* pp. 167-195.

DEM95 de Ridder, H., Blommaert, F. J. J., and Fedorovskaya, E. A., (1995). Naturalness and image quality: Chroma and hue variation in colour images of natural scenes. *Proc. SPIE, Vol. 2411*, San Jose, CA, pp. 51-61.

DIM93 Dimolitsas, S., Corcoran, F. L., and Phipps J. G. Jr, (1993). Impact of transmission delay on ISDN video telephony. *Proc. of Globecom '93 – IEEE Telecommunications Conference*, Houston, TX, pp. 376-379.

DIN99 Dinh H.Q., Walker N., Bong C., Kobayashi A. and Hodges L. F., (1999). Evaluating the importance of multi sensory input on Memory and the sense of Presence in Virtual Environments. *Proc. of IEEE Virtual Reality*, pp. 222-228.

DOG02 Dogan, S., Eminsoy, S., Sadka, A. H., and Kondoz, A. M., (2002) Personalised multimedia services for real-time video over 3G mobile networks. *Third International Conference on 3G Mobile Communication Technologies, 2002. (Conf. Publ. No. 489)*, pp. 366–370.

DUC00 Duchowski, A., Shivashankaraiah, V., Rawis, T., Gramopadhye, A. K., Melloy, B. J., and Kanki, B., (2000). Binocular Eye Tracking in Virtual Reality for Inspection Training. *Proc. of the Eye Tracking Research and Applications Symposium ,* Palm Beach Gardens, Florida, USA, pp. 89-96.

EBI02 Ebina, O., Owada, N., Ohinata, Y., Adachi, K., and Fukushima, M., (2002). Wearable Internet Appliances and Their Applications. *Hitachi Review Vol. 51* (1), pp. 7-11.

ECK93 Eckert M. P., Buchsbaum G., (1993). The significance of eye-movements and image acceleration for colour television image sequences. *Digital Images and Human Vision*, Watson A. B. (Ed.), MIT Press, pp. 89-98.

ELI84 Elias, G., Sherwin, G., and Wise J. (1984). Eye movements while viewing NTSC format television. *SMPTE Psychophysics Subcommittee White Paper.*

ELT02 Elting, C., Zwickel, J., Malaka, R., (2002). Device-Dependent Modality Selection for User-Interfaces – An Empirical Study. *Proc. of ACM IUI'02,* pp. 55-62.

END94 Endo, C., Takuya, A., Haneishi, H., and Miyake, Y., (1994). Analysis of the eye movements and its application to image evaluation. *Proc. Colour Imaging Conference.* Scotdale, AZ, pp. 153-155.

ENG95a Engell-Nielsen, T., and Glenstrup, A. J., (1995). Thesis for the Partial Fulfilment of the Requirements for a Bachelor's Degree in Information Psychology*, Laboratory of Psychology, University of Copenhagen.* Retrieved from University of Copenhagen (July 2004) http://www.diku.dk/~panic/eyegaze

ENG95b __________________ /node16.html - Eye Movements

ENG95c __________________ /node8.html - Present-day Eye-Gaze Tracking Techniques.

ENG95d __________________ /node9.html - Techniques Based on Reflected Light.

ENG95e ________________ /node10.html - A Technique Based on Electric Skin Potential.

ENG95f ________________ /node10.html - Techniques Based onContact Lenses.

ENR96 Enroth-Cugell, C., and Robson, J. G., (1996). The contrast sensitivity of retina ganglion cells of a cat. *Journal of Physiology. Vol. 187,* pp. 517-552.

FAR96 Faraday, P. and Sutcliffe, A., (1999). An Empirical Study of Attending and Comprehending Multimedia Presentations. In *Proc. of ACM Multimedia '96,* Boston, Massachusetts, USA, pp. 265-275.

FAR99 Faraday, P. and Sutcliffe, A., (1999). Authoring Animated Web Pages using Contact Points. In *Proc. of ACM CHI '99,* Pittsburgh, Pennsylvania, USA, pp. 458-465.

FIN80 Findlay, J. M., (1980). The visual stimulus for saccadic eye movements in human observers. *Perception*, *Vol. 9,* pp. 7-20.

FIS03 Fisher, R., Perkins, S. , Walker, A. and Wolfart, E. (2003) Robert Edge Cross Detector. Retrieved from University of Edinburgh (July 2004) http://homepages.inf.ed.ac.uk/rbf/HIPR2/roberts.htm

FOL94 Foley J. M. (1994). Human luminance pattern-vision mechanisms: Masking experiments require a new model. *Journal of the Optical Society of America A, Vol. 11* (6), pp. 1710-1719.

FOX96 Fox, A., Gribble, S.D., Brewer, E.A.., and Amir, E., (1996). Adapting to Network and Client Variability via On-Demand Dynamic Distillation. *Proc. of the Seventh International ACM Conference on Architectural Support for Programming Languages and Operating Systems,* Cambridge, MA, pp. 160-170.

FOX98 Fox, A., Goldberg, I., Gribble, S.D., Lee, D.C., Polito, A. and Brewer, E.A, (1998). Experience With Top Gun Wingman: A proxy-based

Graphical Web Browser for the 3Com PalmPilot. In *Proc. of Middleware '98, Lake District,* England, pp. 407-424.

FRE97 Fredericksen R. E., and Hess R. F., (1997). Temporal detection in human vision: Dependence on stimulus energy. *Journal of the Optical Society of America A Vol. 14* (10), pp. 2557- 2569.

FRE98 Fredericksen R. E., and Hess R. F., (1998). Estimating multiple temporal mechanisms in human vision. *Vision Resolution Vol. 38* (7), pp. 1023-1040.

FUL01 Fulk, M., (2001). Improving Web Browsing on Handheld Devices, In *ACM CHI 2001*, GVU Center & College of Computing, Georgia Institute of Technology, pp. 395-396.

FUK97 Fukuda, K., Wakamiya, N., Murata, M., and Miyahara, H (1997) QoS Mopping between User's Preference and Bandwidth Control for Video Transport. In *Proc. of the 5th International Workshop on QoS (IWQoS)*, New York USA, pp.291-301.

GAL97 Gale, A. (1997). Human response to visual stimuli. *The Perception of Visual Information*, Hendee W. and Wells P., (Eds.), Springer-Verlag, pp. 127-147.

GEE00 Geelhoed, E., Falahee, M, and Latham, K., (2000). Safety and Comfort of Eyeglasses Displays. *Proc. of Handheld and Ubiquitous Computing : Second International Symposium*, HUC 2000, Bristol, pp. 236-247.

GHI98 Ghinea, G., and Thomas, J.P. (1998). QoS Impact on User Perception and Understanding of multimedia Video Clips, *Proc. of ACM Multimedia '98*, Bristol UK, pp. 49- 54.

GHI99 Ghinea G., Thomas, J.P. and Fish, R.S. (1999). Multimedia, Network Protocols and Users - Bridging the Gap, *Proc. of ACM Multimedia '99*, Orlando, Florida, pp. 473-476.

GHI00 Ghinea, G. and Thomas, J.P. (2000). Impact of Protocol Stacks on Quality of Perception, *Proc. IEEE International Conference on Multimedia and Expo*, *Vol. 2*, New York, pp. 847 -850.

GHM01 Ghinea, G., and Magoulas, G., (2001). Quality of Service for Perceptual Considerations: An Integrated Perspective. *IEEE International Conference on Multimedia and Expo*, Tokyo, pp. 752-755.

GHP00 Ghinea, G., (2000). Quality of Perception – An Essential Facet of Multimedia Communications. Submitted for the Degree of Doctor of Philosophy, Department of Computer Science, The University of Reading, UK.

GHT00 Ghinea, G., Thomas, J. P. and Fish, R.S., (2000). Mapping Quality of Perception to Quality of Service: the Case for a Dynamically Reconfigurable Communication System, *Journal of Intelligent Systems, Vol. 10* (5/6), pp. 607-632.

GHT01 Ghinea G. and Thomas J.P., (2001). Crossing the Man-Machine Divide: A Mapping Based on Empirical Results. *Journal of VLSI Signal Processing, Vol. 29* (1/2), pp. 139-147.

GIP93 Gips, J., Olivieri, P. and Tecce, J., (1993). Direct control of the computer through electrodes placed around the eyes. *Fifth International Conference on Human-Computer Interaction*, *(HCI International '93),* Elsevier, Orlando, Florida, pp. 630-635.

GIP96 Gips, J., DiMattia, P., Curran, F.X., Olivieri, C. P., (1996). Using EagleEyes -- an Electrodes Based Device for Controlling the Computer with Your Eyes -- to Help People with Special Needs. *Interdisciplinary Aspects on Computer Helping with Special Needs*, Klausm, J., Auff, E., Kremser, W., Zagler, W., and Oldenbourg, R., (Eds.), Vienna. Available.

GLE95 Glenstrup, A. J., Engell-Nieson, T, (1995). Eye Controlled Media: Present and Future State. Retrieved from University of Copenhagen (July 2004) http://www.diku.dk/users/panic/eyegaze/article.html

GIR89 Girod, B., (1989). The information theoretical significance of spatial and temporal masking in video signals. *Proc. SPIE, Vol. 1077*, Los Angeles, CA, pp. 178-187.

GOL02 Goldberg, J.H., Stimson, M. J., Lewenstein, M., Scott, N., Wichansky A. M., (2002). Eye tracking in web search tasks: design implications. *Proc. of the symposium on ETRA 2002: eye tracking research & applications symposium 2002,* New Orleans, Louisiana, pp. 51-58.

GRA93 Gray, J., and Reuter, A., (1993). Transaction Processing: Concepts and Techniques. Morgan Kaufman.

GRI98 Gringeri, S., Khasnabish, G., Lewis, A., Shuaib, K., Egorov, R., and Bash, B. (1998). Transmission of MPEG-2 video streams over ATM. *Proc. IEEE Multimedia*, *Vol. 5* (1), pp. 58-71.

GUL03 Gulliver S.R. and Ghinea G. (2003). How Level and Type of Deafness Affects User Perception of Multimedia Video Clips, *Universal Access in the Information Society*, *Vol. 2*, (4), pp. 374-386.

GUL04 Gulliver, S. R. and Ghinea, G. (2004). Starts in their Eyes: What Eye-Tracking Reveal about Multimedia Perceptual Quality. *IEEE Transaction on System, Man and Cybernetics, Part A, Vol. 34* (4), pp. 472-482.

HAD98 Hardman, V., Sasse, M. A., and Kouvelas, I., (1998). Successful multi-party audio communications over the Internet. *Communication of the ACM*, *Vol. 41* (5), pp. 74-80.

HAR03 Harroud, H., Ahmed, M., and Karmouch, A., (2003). Policy-driven personalized multimedia services for mobile users. *IEEE Transactions on Mobile Computing, Vol. 2* (1), pp. 16 – 24.

HAY02 Hayhoe, M. M., Ballard, D. H., Triesch, J., Aivar, P., Sullivan, B., (2002). Vision in natural and virtual environments. *Proceedings of the symposium on ETRA 2002: eye tracking research & applications symposium 2002*, New Orleans, Louisiana, pp. 7-13.

HAT56 Hartline, H.K., Wagner, H. G. and Ratliff, F. (1956). Inhibition in the eye of Limulus. *Journal of General Physiology*, *Vol. 39,* pp. 651-673.

HAS93 Hasser, C., and Weisenberger, J. M. (1993) Preliminary evaluation of a shape memory alloy tactile feedback display. In *Proc. Symposium on Haptic Interfaces for Virtual Environments and Teleoperator Systems*, Kazerooni, H., Adelstein, B.D., Colgate, J.E., (Eds.), ASME Winter Annual Meeting, New Orleans, LA, pp. 73-80.

HEI62 Heilig, M.L. (1962) Sensorama Simulator. U.S.Patent 3,050,870.

HEI92 Heilig, M.L. (1992) El Cine del Futuro: The cinema of the future. *Presence, Vol. 1,* (3), pp. 279-294.

HER78 Hering, E. (1878) Zur lehure vom Lichtsinne. Carl Gerolds & Sohn, Vienna, Austria.

HES99 Hess C. K., and Campbell R. H. (1999) Media Streaming Protocol: An Adaptive Protocol for the Delivery of Audio and Video over the Internet. *Proc. of IEEE ICMCS'99*, Florence, Italy, pp. 903-907.

HIL94 Hillstrom, A. P. and Yantis, S., (1994) Visual motion and attentional capture. *Perception and Psychophysics, Vol. 55,* pp. 399-411.

HOF78 Hoffman, J.E., (1978). Search through a sequentially presented visual display. *Perception and Phychophysics Vol. 23,* pp. 1-11.

HOL97 Hollier, M. P. and Voelcker, R. M. (1997) Towards a multimodal perceptual model, *BT technological Journal, Vol. 15* (4), pp. 162-171.

HOR96 Horita, Y., Katayama, M., Murai, T., and Miyahara, M. (1996) Objective picture quality scale for video coding. *Proc. ICIP, Vol. 3,* Lausanne, Switzerland, pp. 319 – 322.

HOW95 Howe, R.D., Peine, W.J., Kontarinis D.A., and Son, J.S. (1995) Remote palpation technology. *IEEE Engineering in Medicine and Biology, Vol. 14* (3), pp. 318-323.

HUB70 Hubel, D. H. and Wiesel, T. N., (1970) The period of susceptibility to the physiological effects of unilateral eye closure in kittens. *Journal of Physiology, Vol. 206,* pp. 436.

HUT89 Hutchinson, T. (1989). Human-computer interaction using eye-gaze input. *IEEE Transactions on Systems, Man and Cybernetics, Vol. 19*, pp. 1527-1534.

IAI93 Iai, S., Kurita, T., and Kitawaki, N. (1993) Quality requirements for multimedia communication services and terminal – interaction of speech and video delays. *Proc. of Globecom '93 – IEEE Telecommunications Conference,* Houston, TX, pp. 394-398.

INO93 Ino, S., Shimizu, S., Odagawa, T., Sato, M., Takahashi, M., Izumi, T., and Ifukube, T. (1993) A tactile display for presenting quality of materials by changing the temperature of skin surface. *Proc.. Second IEEE*

International Workshop on Robot and Human Communication, Tokyo, pp. 220-224.

ISO00 Isokoski, P, (2000). Text input methods for eye trackers using off-screen targets. *Proc. of the symposium on Eye tracking research and applications 2000,* Palm Beach Gardens, Florida, United States, pp. 15-21.

ITU50 ITU-R BT.500-7 : *Methodology for the subjective assessment of television pictures*

ITU80 ITU-T P.800 : *Methods for subjective determination of transmission quality*

JAC91 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, Vol. 9* (3), pp. 152-169.

JAC95 Jacob, R. J. K., (1995). Eye tracking in advanced interface design. *Advanced Interface Design and Virtual Environments*, Barfield W. and Furness T. (Ed..), Oxford University Press, Oxford, pp. 258-288.

JON99 Jones, M., Marsden, G., Mohd-Nasir, N., Boone, K. and Buchanan, G., (1999). Improving Web interaction on Small Displays. *Proc.. of 8th International WWW Conference,* pp. 51-59.

JUD89 Juday R. D. and Fisher T. E., (1989) Geometric transformations for video compression and simulation. *Society for information Display, Vol.. 39,* pp 420-423.

JUS76 Just, M. A. and Carpenter, P. A., (1976) Eye Fixations and Cognitive Processes. *Cognitive Psychology*, *Vol.. 8,* pp. 441-480.

KAH73 Kahneman D., (1973) Attention and Effort. *Englewood Cliffs, NJ: Prentice-Hall,* 1973.

KAU69 Kaufman, L., Richards, W., (1969) Spontaneous fixation tendencies for visual forms. *Perception and Psychophysics Vol. 5*, pp. 85-88.

KAW95 Kawalek, J. A. (1995) User perspective for QoS Management. *Proc. of the QoS Workshop aligned with the 3rd Internations Conference on Intelligence in Broadband Services and Network (IS&N 95)*, Crete, Greece.

KAZ03 Kazasis, F.G., Moumoutzis, N. Pappas, N., Karanastasi, A, and Christodoulakis, S., (2003). Designing Ubiquitous Personalized TV-Anytime Services. *Proc. of CAISE'03 Workshops*, Eder, J., Mittermeir, R., Pernici, B. (Eds.), University of Maribor Press, Slovenia, pp. 136-149.

KEN97 Kennedy, R.S., Lanham, D.S, Drexler, J.M., Massey, C.J, and Lilienthal, M.G., (1997). A comparison of cybersickness incidences, symptom profiles, measurement techniques, and suggestions for further research. In *Teleoperator & Virtual Environment, Vol. 6* (6), pp. 638-644.

KIE97 Kies J. K., Williges, R.C. and Rosson, M. B. (1997) Evaluating desktop video conferencing for distance learning. *Computers and Education, Vol. 28,* pp. 79-91.

KIJ97 Kijima, R., and Ojika, T., (1997). Transition Between Virtual Environment and Workstation Environment with Projective Head Mounted Display. *Proc. of IEEE Virtual Reality Annual International Symposium*, Albuquerque, New Mexico: IEEE Computer Society Press, pp. 130-137.

KIM01 Kim, L., and Albers, M.J., (2001). Web Design Issues when Searching for Information in a Small Screen Display. *Proc. ACM SIGDOC'01,* University of Memphis, pp. 193-200.

KIS92 Kistler, J.J., and Satyanareyanan, M., February (1992). Disconnected Operation in the Coda File System. *ACM Transactions on Computer Systems Vol. 10* (1), pp. 3-25.

KLE93 Klein, S. A., (1993). Image quality and image compression: A psychophysicist's viewpoint. In Digital Images and Human Vision, Watson A. B. (Ed.), MIT Press, pp. 73-88.

KOL95 Kolasinki, E.M., (1995). Simulator sickness in virtual environments. Retireved from Technical Report 1027, U.S. Army Research Institute for the Behaviour and Social Sciences, Alexandria, (July 2004) http://www.cyberedge.com/3a5_2.html

KOL90 Kolb, B., and Whishaw, I. (1990). Fundamentals of Human Neuropsychology. W.H. Freeman and Co., New York..

KON95 Kontarinis, D. A., and Howe, R. D., (1995) Tactile display of vibratory information in teleoperation and virtual environments. *Presence, Vol. 4* (4), pp. 387-402.

KOO98 Koodli, R., and Krishna, C.M., (1998). A loss model for sorting QoS in multimedia applications. *Proc.. of ISCA CATA-98: Thirteenth International Conference on Computers and Their Applications*, ISCA, Cary, NC, USA, pp. 234-237.

KOR96 Kortum, P. T., and Geisler, W. S., (1996). Implementing of a foveated image-coding system for bandwidth reduction of video images. *SPIE Proc. Human Vision and Electronic Imaging, Vol. 2657*, pp. 350-360.

KOW90 Kowler, E. (1990). The role of visual and cognitive processes in the control of eye movement. In *Eye movements and their Role in Visual and Cognitive Processes*, Kowler, E. (Ed.), Elsevier Science Publishers.

KRA03 Krasic, C., Walpole, J., and Feng W., (2003). Quality-Adaptive Media Streaming by Priority Drop. *Proc. NOSSDAV'03,* Monterey, California, USA, pp. 307–310.

KRL04 Krüger, N., Lappe, M., and Wörgötter, F., (2004). Biologically Motivated Multi-modal Processing of Visual Primitives. *Interdisciplinary Journal of Artificial Intelligence the Simulation of Behavious, AISB Journal, Vol.1* (5).

KRW04 Krüger, N., Felsberg, M., and Wörgötter, F., (2004). Processing Multi-modal Primitives from Image Sequences. *Proc. of the Fourth International ICSE Symposium on Engineering of Intelligent Systems*, EIS'2004, Island of Madeira, Portugal.

KUF53 Kuffler, S.W., (1953). Discharge patterns and functional organization of the mammalian retina. *Journal of Neurophysiolology*, *Vol. 16,* pp. 37-68.

LAN97 Lantz, E., (1997). Future Directions in Visual Display Systems. *Computer Graphics, Vol. 31* (2), pp. 38-45.

LAN00 Lankford, C., (2000). Effective Eye-gaze Input into Windows. In *Proc. of the Eye Tracking Research and Applications Symposium*, Palm Beach Gardens, Florida, USA, pp. 23 -27.

LAP00 Lappin, J. S., and Craft, W. D., (2000) Foundations of Spatial Vision: From Retinal Images to Perceived Shapes. *Psychological Review, Vol.* 107 (1), pp. 6-38.

LIN96 Lindh, P. and van den Branden Lambrecht, C. J., (1996). Efficient Spatio-Temporal Decomposition for Perceptual Processing of Video Sequences. *Proc. of the International Conference on Image Processing, Vol. 3,* Lausanne, Switzerland, pp. 331-334.

LOD96 Lodge, N., (1996). An introduction to advanced subjective assessment methods and the work of the MOSAIC consortium. *MOSAIC Handbook*, pp. 63-78.

LOS94 Losada, M. A., and Mullen, K. T., (1994). The spatial tuning of chromatic mechanisms identified by simultaneous masking. *Vision Resolution, Vol. 34* (3), pp. 331-341.

LOS00 Loschky, L. C., and McConkie, G. W., (2000). User performance with Gaze-Contingent Multiresolutional Displays. *Proc. of ACM Eye Tracking Research and Applications Symposium (ETRA) 2000,* Palm Beach Gardens, FL, pp. 97-103.

LUK82 Lukas, F. X. J., Budbikis, Z. L., (1982). Picture quality prediction based on a visual model. *IEEE Transactions on Communications*, *Vol. 30* (7), pp. 1679–1692.

MAC67 Mackworth, J. F., and Morandi, A. J., (1967). The gaze selects informative details within pictures. *Perception and Psychophysics, Vol. 2,* pp. 547-552.

MAC70 Mackworth, J. F., and Bruner, J. S., (1970). How adults and children search and recognize pictures. *Human Development*, *Vol. 13,* pp. 149-177.

MAE96 Maeder, A., Diederich, J., and Niebur, E., (1996). Limiting human perception for image sequences. *Proc. SPIE, Vol. 2657,* San Jose, CA., pp. 330-337.

MAH03 Mahanti, A., Eager, D. L., Vernon, M. K., and Sundaram-Stukel, D. J., (2003). Scalable on-demand media streaming with packet loss recovery, *IEEE/ACM Transactions on Networking, Vol. 11* (2), pp. 195-209.

MAJ02 Majaranta P., and Räihä K., (2002). Twenty years of eye typing: systems and design issues. *Proc. of the symposium on Eye Tracking Research and Applications Symposium (ETRA) 2002*, New Orleans, Louisiana, pp. 15-22.

MAN74 Mannos, J. L., and Sakrison D. J. (1974) The effects of a visual fidelity criterion on the encoding of images. IEEE *Transactions on Information Theory, Vol. 20* (4), pp. 525-536.

MAR96 Martens, J. B., and Kayargadde, V., (1996). Image quality prediction in a multidimensional perceptual space. *Proc. ICIP, Vol. 1,* Lausanne, Switzerland, pp. 877-880.

MAS01 Masry, M., Hemami, S.S, Osberger, W.M. and Rohaly, A.M., (2001). Subjective quality evaluation of low-bit-rate video. *Human vision and electronic imaging VI – Proc. of the SPIE*, Rogowitz, B.E. and Pappas, T.N. (Eds.), SPIE, Bellingham, WA, USA, pp. 102-113.

MAS95 Massimino, M. J. and Sheridan, T. B., (1995). Teleoperator performance with varying force and visual feeback. *Human Factors,* pp. 145-157.

MAT91 Matsui, T., Hirahara, S., (1991). A new human vision system model for spatio-temporal image signal. *Proc. SPIE, Vol. 1453*, San Jose, CA, pp. 282-289.

MAY97 Mayer, R. E., (1997). Multimedia Learning: Are We Asking the Right Questions?, *Educational Psychologist, Vol.. 32* (1), pp. 1–19.

MCC75 McConkie, G. W. and Rayner, K., (1975). The span of the effective stimulus during a fixation in reading. *Perception and Phychophysics, Vol.. 17,* pp. 578-586.

MEY99 Meyer, M., and Barr, A., (1999). ALCOVE: Design and Implementation of an Object-Centric Virtual Environment. *Proc.s of IEEE Virtual Reality*, Houston, Texas: IEEE Computer Society Press, pp.46-52.

MIN96 Minsky, M., and Lederman,, S. J., (1996). Simulated Haptic Textures: Roughness. *Symposium on Haptic Interfaces for Virtual Environment and*

Teleoperator Systems, ASME International Mechanical Engineering Congress and Exposition, Proceedings of the ASME Dynamic Systems and Control Division, Vol.. 58, Atlanta, GA, pp. 451-458.

MON92 Monkman G. J., (1992) Electrorheological Tactile Display, *Presence*, Vol. 1, (2), MIT Press.

MOR93 Mordkoff, J. T., and Yantis, S., (1993) Dividing attention between color and shape: Evidence of coactivation. *Perception and Psychophysics, Vol.. 53*, pp. 357-366.

MUL93 Müller, P. U., Cavegn, D., d'Ydewalle, G. and Groner, R., (1993). A comparison of a new limbus tracker, corneal reflection technique, purkinje eye tracking and electro-oculography, *In* `Perception and Cognition', Elsevier Science Publishers, d'Ydewalle G., and Rensbergen J. V., (Eds.), pp. 393-401.

MULL01 Mullin, J., Smallwood, L., Watson, A.., and Wilson, G. M., (2001). New Techniques for Assessing Audio and Video Quality in Real-Time Interactive Communication. *Proc. of IHM-HCI 2001*, Vanderdonckt, J., Blandford, A. and Derycke, A. (Eds.), Lille, France, pp. 221-222.

MUNN32 Munn, L., and Geil, G., (1932). A note on peripheral form discrimination, *Journal of General Psychology, Vol.5*, pp. 78-78.

NAK02 Nakamoto, T., and Hiramatsu, H., (2002). Study of odour recorder and dynamic change in odour using QCM sensors and neural networks. *Sensors and Actuators. B: Chemical, Vol.. 85,* (3), pp. 263-269.

NEE78 Needham, R.M. and Schroeder, M.D., (1978). Using Encryption for Authentication in Large Networks of Computers. *Communications of the ACM, Vol.. 21* (12).

NIE94 Niebur, E. and Koch, C., (1994). A model for the neuronal implementation of selective visual attention based on temporal correlation among neurons. *Journal of Computational Neuroscience, Vol.. 1* (1), pp. 141-158.

NIE97 Niebur, E. and Koch C., (1997). Computational architectures for attention. In *The Attentive Brain*. Parasuraman, R. (Ed.), MIT Press.

NIE00 Nielsen, J., (2000). Jakob Nielsen's Alertbox 2000: Eyetracking Study of Reading. Retrieved from useit.com (July 2004) http://www.useit.com/alertbox/20000514.html

OSB97 Osberger, W., Maeder, A. J., and McLean, D., (1997). A Computational model of the human visual system for image quality assessment. *Proc.. DICTA-97*, Auckland, New Zealand, pp. 337-342.

OSB98 Osberber, W., Maeder, A. J., and Bergmann, N. W., (1998). A technique for image quality assessment based on a human visual system model. *9th European Signal Processing Conference (EUSIPO-98)*, Rhodes Greece, pp. 1049-1052.

OSM98a Osberger, W., and Maeder, A. J., (1998). Automatic identification of perceptually important regions in an image using a model of the human visual system. *14th International Conference on Pattern Recognition,* Brisbane, Australia.

OSM98b Osberger, W., Maeder, A. J., and Bergmann, N., (1998). A perceptually based quantization technique for MPEG encoding. *SPIE Human Vision and Electronic Imaging III, Vol.* 3299.

PAB02 Pablo Research Group, (2002). Intelligent Information Spaces. Retrieved from Department of Computer Science, University of Illinois at Urbana-

Champaign (July 2004) http://www-pablo.cs.uiuc.edu/Project/SmartSpaces/SmartSpaceOverview.htm

PAR00 Parkhurst, D., Culurciello, E., and Niebur, E., (2000). Evaluating Variable Resolution Displays with Visual Search: Task Performance and Eye Movements. *Proc. of the Eye Tracking Research and Applications Symposium,* Palm Beach Gardens, Florida, USA, pp. 105 -109.

PAR02 Parkhurst, D. J., and Niebur, E., (2002). Variable resolution display: A theoretical, practical and behavioural evaluation. *Human Factors, Vol. 44* (4), pp. 611-629.

PAS00 Partala, T., Jokiniemi, M., Surakk, V., Pupillary, (2000). Responses to emotionally provocative stimuli. *Proc. of the symposium on Eye tracking research and applications 2000*, Palm Beach Gardens, Florida, United States, pp. 123-128.

PEL90 Peli, E., (1990). Contrast in complex images. *JOSA, Vol.* 7 (10), pp. 203-204.

PEL00 Pelz, J.B., Canosa, R., Babcock, J., (2000). Extended Tasks Elicit Complex Eye Movement Patterns. *Proc. of the symposium on Eye tracking research and applications 2000*, Palm Beach Gardens, Florida, USA, pp. 37-43.

PER98 Perkins, C., Hodson, O., and Hardman, V. (1998). A survey of packet-loss recovery techniques for streaming audio. *IEEE Network Magazine*, September/October.

POM93 Pomerleau, D. and Baluja, S., (1993). Non-Intrusive Gaze Tracking Using Artificial Neural Networks, *AAAI Fall Symposium on Machine Learning in Computer Vision*, Raleigh, NC, pp. 153-156.

PRO99 Procter, R., Hartswood, M., McKinlay, A. and Gallacher, S., (1999). An investigation of the influence of network quality of service on the effectiveness of multimedia communication. *Proc. of the international ACM SIGGROUP conference on supporting group work*. ACM. New York, NY, USA, pp. 160-168.

QUA02 Quaglia D., and De Martin J. C., (2002). Delivery of MPEG Video Streams with Constant Perceptual Quality of Service. *Proc. of IEEE International Conference on Multimedia and Expo (ICME), Vol. 2,* Lausanne, Switzerland, pp. 85-88.

RAD94 Radhika R. R., (1994). Networking constraints in multimedia conferencing and the role of ATM networks. *AT&T Technical Journal*, July/August 1994.

RAY98 Rayner, K., (1998). Eye movements in reading and information processing: 20 years of research. *Pychological Bulletin*, *Vol.. 124* (3), pp. 372-422.

RAZ98 Razdan, R., and Kielar, A., (1998). Eye Tracking for Man/ Machine Interfaces*, Sensors.*

REA81 Read, D., (1981). Solving deductive-reasoning problems after unilateral temporal lobectomy. *Brain and Language*, *Vol.. 12,* pp.116-127.

REG95 Regan, E.C., (1995). An investigation into nausea and other side-effects of head-coupled immersive virtual reality. *Virtual Reality Vol.. 1* (1), pp. 17-32.

REI01 Reingold, E. M., Loschky, L. C., Stampe, D. M., and Shen, J., (2001). An assessment of a Live-video gaze contingent variable resolution display. In *Proceedings of Human-Computer Interaction,* Smith, M.J, Salvendy, G., Harris, D. and Koubek, R. J., (Eds.), Lawrence Earlbaum Associates, Mahwan, N.J.L USA, pp. 1338-1342.

REI02 Reingold, E., Loschky, L. C., (2002). Reduced saliency of peripheral targets in gaze-contingent multi-resolutional displays: blended versus sharp boundary windows. *Proc. of the symposium on ETRA 2002: eye tracking research and applications symposium 2002,* New Orleans, Louisiana, pp. 89-93.

RIM98 Rimmel, A. M., Hollier, M. P., and Voelcker, R. M., (1998). The influence of cross-modal interaction on audio-visual speech quality perception, *Presented at the 105th AES Convention*, San Francisco, CA: Available.

RIM99 Rimmell, A. M., and Hollier M. P., (1999). The significance of cross-modal interaction in audio-visual quality perception, *Multimedia Signal Processing,* IEEE Signal Processing Society 1999 Workshop on Multimedia Signal Processing, pp. 509-514.

ROB94 Robson, R., (1994). *Experiment, Design and Statistics in Psychology* (3rd ed.) Penguin.

ROD65 Rodieck, R. W., and Stone, J. S., (1965). Analysis of receptive fields of cat retinal ganglion cells. Journal *of Neurophysiology, Vol.. 28,* pp. 965-980.

ROU92 Roufs, J. A. J., (1992). Perceptual image quality: Concept and measurement. *Philips Journal of Resolution, Vol.. 47* (1), pp. 35-62.

ROY94 Roy, R. R., (1994). Networking constraints in multimedia conferencing and the role of ATM networks. *AT&T Technical Journal,* July/August.

ROY99 Royer, E.M., Toh, C.K., (1999). A Review of Current Routing Protocols for Ad Hoc Mobile Wireless Networks. *IEEE Personal Communications Vol.. 6* (2), pp. 46-55.

RYA01 Ryans, M. A., Homer, M. L., Zhou, H., Manatt, K, and Manfreda, A., (2001). Toward a Second Generation Electronic Nose at JPL: Sensing Film Organisation Studies. *Proc. International Conference on Environmental Systems'01*. Available.

SAI79 Saida, S., and Ikeda M., (1979). Useful visual field size for pattern perception. *Perception and Psychophysics, Vol. 25*, pp. 119-125.

SAL66 Salapatel, P., and Kessen, W., (1966). Visual scanning of triangles in the human newborn, *Journal of Experimental Child Psychology*, Vol. 3 (2), pp. 155- 167.

SAL99 Salvucci, D. D., (1999). Inferring Intent in Eye-based Interfaces: Tracing Eye Movements with Process Models. *Proc. of ACM CHI '99*, Pittsburgh, Pennsylvania, USA, pp. 254-261.

SAN70 Sanders, A. F., (1970). Some aspects of the selective process in the functional visual field, *Ergonomics, Vol. 13*, pp 101-117.

SAT89 Satyanarayanan, M., (1989). A Survey of Distributed File Systems. In *Annual Review of Computer Science,* Traub, J.F, Grosz, B., Lambson, B., Nilsson, N.J. (Eds.), Annual Reviews Inc., pp. 73-104.

SAT93 Satoru, I., Takaaki, K., and Nobuhiko, T., (1993). Quality requirements for multimedia communication services and terminal – interaction of speech and video delays. *Proc. of Globecom '93 – IEEE Telecommunications Conference*, Houston, TX, pp. 394-398.

SAT01 Satyanarayanan, M., (2001). Pervasive Computing: Vision and Challenges. *IEEE Personal Communications Vol.. 6* (8), Carnegie Mellon University. pp. 10-17.

SCI00 Scialfa, C. T., McPhee, L., and Ho, G., (2000). The Effects of a Simulated Cellular Phone Conversation on Search for Traffic Signs in an

Elderly Sample. *Proc. of the Eye Tracking Research and Applications Symposium,* Palm Beach Gardens, Florida, USA, pp. 45-50.

SCH56 Schade, O. H., (1956). Optical and photoelectric analogue of the eye. *Journal of the Optical Society of America A., Vol.* 46 (9), pp 721-739.

SCO92 Scott, D., and Findlay, J. M., (1992). Visual search, eye movements and display units. *Human factors report*, University of Durham, South Road, Durham DH1 3LE, UK.

SEN92 Senders, J., (1992). Distribution of attention in static and dynamic scenes. P*roc. SPIE Vol. 3016,* San Jose, pp. 186-194.

SHI89 Shioiri, S. and Ikeda, M., (1989). Useful resolution for picture perception as a function of eccentricity. *Perception,* 18, pp. 347-361.

SIL93 Silsbee, P. L., Bovik, A. C., and Chen, D., (1993). Visual Pattern image sequence coding. *IEEE Transactions on Circuit and Systems for Video Technology*, Vol. 3, pp. 291-301.

SIM90 Simoncelli, E. P., and Adelson, E.H., (1990). Non-separable extensions of quadrature mirror filters to multiple dimensions. *Proc. IEEE Special Issue on Multi-dimensional Signal Processing, Vol. 78* (4), pp. 652-664.

SIM92 Simoncelli, E. P., Freeman, W. T., Adelson, E. H. and Heeger, D. J., (1992). Shiftable multi-scale transforms". *IEEE Trans. Information Theory Vol.. 38* (2), pp. 587-607.

SIM95 Simoncelli E. P., Freeman W. T., (1995). The steerable pyramid: A flexiable architecture for multi-scale derivative computation. *Proc. ICIP,* Washington, DC, pp. 444-447.

SOD02 Sodhi, M., Reimer, B., Cohen, J.L, Vastenburg, E., Kaars, R., and Kirschenbaum, S., (2002). Onroad driver eye movement tracking using head-mounted devices. *Proc. of the symposium on ETRA 2002: eye tracking research & applications symposium 2002,* New Orleans, Louisiana, pp. 61-68.

STE90 Steinmetz, R., (1990). Synchronisation properties in multimedia systems. *IEEE Journal Select. Areas Communication, Vol.. 8* (3), pp. 401-412.

STE92 Steinmetz R., (1992). Multimedia Synchronisation techniques experiences based on different systems structures. *Proc. IEEE Multimedia Workshop '92*, Monterey, CA.

STE96 Steinmetz, R., (1996). Human Perception of Jitter and Media Synchronisation. *IEEE Journal on Selected Areas in Communications*, *Vol.. 14* (1), pp. 61-72.

STE97 Stelmach, L. B., Campsall, J. M., and Herdman, C. M., (1997). Attentional and Occular Movements. *Journal of Experimental Psychology: Human Perception and Performance*, *Vol. 23*, pp. 823-844.

STB96 Steinmetz, R. Blakowski, G., (1996). A media Synchronisation survey: Reference model, specification and case studies. *IEEE Journal of Selected Areas of Communications Vol.. 14* (1) pp 1-4.

STT91 Stelmach, L. B., Tam, W. J. and Hearty, P. J., (1991). Static and dynamic spatial resolution in image coding: An investigation of eye movements. *Proc. SPIE, Vol.. 1453,* San Jose, CA, pp. 147-152.

STT94 Stelmach, L.B., Tam, W. J., (1994). Processing image based on eye movements. *Proc. SPIE on eye movements, Vol. 2179,* San Jose, CA, pp. 90-98.

SWA93 Swartz, M. and Wallace, D., (1993). Effects of frame rate and resolution reduction of human performance. *Proc. of IS&T's 46th Annual Conference*, Munich, Germany.

SWI88 Switkes, E., Bradley, A., and De Valois, K. K., (1988). Contrast dependence and mechanisms of masking interactions among chromatic and luminance gratings." *Journal of the Optical Society of America A, Vol.. 5* (7), pp. 1149-1162.

TAN98 Tan, K. T., Ghanbari, M., and Pearson, D. E., (1998). An objective measurement tool for MPEG video quality. *Signal Processing, Vol.. 70* (3), pp. 279-294.

TEO94 Teo, P. C. and Heeger, D. J., (1994). Perceptual Image Distortion. *Human Vision, Visual Processing and Digital Display V, IS&T / SPIE's Syposium on Electronic Imaging: Science & Technology, Vol. 2179*, San Jose, CA, pp. 127-141.

TRE86 Tresman, A., (1986). Features and objects in visual processing", *Scientific American, Vol.. 255* (5), pp 106-115.

TUR84 Turner, J. A., (1984). Evaluation of an eye-slaved area-of interest display for tactical combat simulation. *6th Interservice / Industry Training Equipment Conference and Exhibition,* pp. 75-86.

TVA03 TV-Anytime, (2003). TV Anytime Forum Website, retrieved from TV-Anytime.org (July 2004) http://www.tv-anytime.org

MUR04 Murphy, T. E., Webster, R. J., (3rd), and Okamura, A. M., (2004). Design and Performance of a Two-Dimensional Tactile Slip Display, *EuroHaptics 2004*, Technische Universität München, Munich, Germany.

TOW93 Towsley, D., (1993). Providing quality of service in packet switched networks. *Performance of Computer Communications*, Donatiello L., and Nelson R. (Eds.), Springer, Berlin Heidelberg, New York, pp. 560-586.

VAF96 van den Branden Lambrecht, C. J., Farrell, J. E., (1996). Perceptual quality metric for digitally coded color images. *Proc. of the VIII European Signal Processing Conference EUSIPCO,* Trieste, Italy, pp. 1175-1178.

VAN96 van den Bradnden Lambrecht, C. J., (1996). Colour moving pictures quality metric. *Proc. ICIP, Vol.. 1,* Lausanne, Switzerland, pp 885-888.

VAV96 van den Branden Lambrecht, C. J., and Verscheure, O., (1996). Perceptual Quality Measure using a Spatio-Temporal Model of the Human Visual System. *Proc. of the SPIE, Vol. 2668,* San Jose, CA, pp. 450-461.

VEB96 Verscheure, O., Basso, A., El-Maliki, M., and Hubaux J-P., (1996). Perceptual Bit Allocation for MPEG-2 Video Coding. *Proc. of the International Conference on Image Processing*, Lausanne, Switzerland.

VEH96 Verscheure, O., and Hubaux, J. P., (1996). Perceptual Video Quality and Activity Metrics: Optimization of Video Services based on MPEG-2 Encoding, *COST 237 Workshop on Multimedia Telecommunications and Applications*, Barcelona.

VEV97 Verscheure, O., and van den Branden Lambrecht, C. J., (1997). Adaptive Quantization using a Perceptual Visibility Predictor, *International Conference on Image Processing (ICIP),* Santa Barbara, pp. 298-302.

VIR95 Virtanen, M.T., Gleiss, N. and Goldstein, M. (1995). On the use of evaluative category scales in telecommunications. *Proc. Human Factors in Telecommunications '95.*

WAN01 Wang, Y., Claypool, M., and Zuo, Z. (2001). An empirical study of RealVideo performance across the internet. *Proc. of the First ACM SIGCOMM Workshop on Internet Measurement.* ACM Press, New York, NY, USA, pp. 295-309.

WAR97 Ward, A., Jones, A., and Hopper, A., (1997). A New Location Techniques for the Active Office. *IEEE Personal Communications, Vol. 4* (5), pp. 42-47.

WAS96 Watson, A., and Sasse, M, A., (1996). Evaluating audio and video quality in low cost multimedia conferencing systems, *Interacting with Computers, Vol. 8* (3), pp 255-275.

WAS97 Watson, A., and Sasse, M.A., (1997). Multimedia conferencing via multicasting: Determining the quality of service required by the end user. *Proc. of AVSPN '97*, Aberdeen, Scotland, pp. 189-194

WAS00 Watson, A., and Sasse, M.A., (2000). The Good, the Bad, and the Muffled: The Impact of Different Degradations on Internet Speech. *Proc. of the 8th ACM International Conference on Multimedia*, Marina Del Rey, CA, pp. 269-302.

WAT87 Watson, A. B., (1987). The cortex transform: Rapid computation of simulated neural images. *Computer Vision, Graphics, and Image Processing Vol. 39* (3), pp. 311-327.

WAT90 Watson, A. B., (1990). Perceptual-components architecture for digital video. *Journal of Optical Society America. Vol. 7* (10), pp. 1945-1954.

WAT94 Watson, A. B., (1994). Perceptual optimisation of DCT colour quantization, matrices, *IEEE International Conference on Image Processing, Vol. 1,* pp. 100-104.

WAT95 Watson, A. B., (1995). Image data compression having minimum perceptual error. US Patent 5,426,512.

WAT97 Watson, A. B., (1997). Image data compression having minimum error. US Patent 5,629,780.

WAT98 Watson, A. B., (1998). Toward a perceptual video metric." *Proc. SPIE, Vol. 3299*, San Jose, CA, pp. 139-147.

WAT96 Watson, B. A., Walker, N., Hodges, L. F., and Warden, A., (1996). Managing level of detail through peripheral degradation: Effects on search performance with a head-mounted display. *ACM Transactions on Computer-Human Interaction, Vol. 4* (4), pp 323-346.

WAY97 Watson, A. B., Yang, G. Y., Solomon, J. A. and Villasenor, J., (1997). Visibility of wavelet quantization noise. *IEEE Transactions on Image Processing, Vol. 6* (8), pp. 1164-1175.

WEI98 Weiser, M., (1998). The Future of Ubiquitous Computing On Campus. *Communications of the ACM, Vol. 41*, pp. 41-42.

WEI90 Weiman, C. F. R., (1990). Video Compression via lop polar mapping, *SPIE Proceedings: Real Time Image Processing II*, Vol. 1295, pp. 266-277.

WES97 Westen, S. J. P., Lagendijk, R. L., and Biemond, J., (1997). Spatio-temporal model of human vision for digital video compression. *Proc. Visual Communications and Image Processing, Vol. 3016,* SPIE - The Int. Society for Optical Engineering, Bellingham.

WIJ96 Wijesekera, D., and Stivastava, J., (1996). Quality of Service (QoS) Metrics for Continuous Media. *Multimedia Tools Applications, Vol. 3* (1), pp 127-166.

WIJ99 Wijesekera, D., Srivastava, J., Nerode, A., and Foresti M, (1999). Experimental Evaluation of loss perception in continuous media, *Multimedia Systems, Vol. 7*, pp. 486-499.

WIL98 Wilson, A., and Sasse, M., (1998). Measuring perceived quality of speech and video in multimedia conferencing applications. *Proc. of ACM Multimedia,* Bristol, UK, pp. 55-60.

WIL00a Wilson, G.M., and Sasse, M.A., (2000). Listen to your heart rate: counting the cost of media quality. *Affective Interactions Towards a New Generation of Computer Interfaces.* Paiva A.M. (Ed.), Springer, Berlin, DE, pp. 9-20.

WIL00b Wilson, G.M., and Sasse, M. A., (2000). Do Users Always Know What's Good For Them? Utilising Physiological Responses to Assess Media Quality. *Proc. of HCI 2000: People and Computers XIV - Usability or Else! Springer,* McDonald S., Waern Y. and Cockton G. (Ed.), pp. 327-339, Sunderland, UK

WIK02 Wikstrand, G., and Eriksson, S., (2002). Football Animation for Mobile Phones, *Proc. of NordiCHI,* pp. 255-258.

WIK03 Wilkstrand, G., (2003). Improving user comprehension and entertainment in wireless streaming media, *Introducing Cognitive Quality of Service*, Department of Computer Science, Umeå, Sweden.

WIN98 Winkler S., (1998). A perceptual distortion metric for digital colour images. *Proc. ICIP, Vol. 3,* Chicago, pp. 399-403.

WIN99 Winkler S. (1999). A perceptual distortion metric for digital colour video. *Proc. SPIE, Vol.. 3644,* San Jose, CA, pp. 175.

WIN01 Winkler, S., (2001). Visual fidelity and perceived quality: toward comprehensive metrics. *Human vision and electronic imaging VI – Proc. of SPIE*, Rogowitz B. E. and Pappas T.N. (Eds.), Bellingham, WA, USA. pp. 114-125.

WOO02 Wooding, D. S., (2002). Fixation Maps: Quantifying Eye-movement Traces, *Proc. of the symposium on ETRA 2002: eye tracking research & applications symposium 2002,* New Orleans, Louisiana, pp. 31 – 36.

XYB03 Xybernaut, (2003). Xybernaut Corporation. Retrieved from Xybernaut.com (July 2004) http://www.xybernaut.com

YAR67 Yarbus, A. L., (1967). Eye movement and vision, trans. B. Haigh. Plenum Press, New York.

YEN98 Yendrikhovski, S. N., Blommaert, F. J. J., and de Ridder, H., (1998). Perceptual optimal colour reproduction. *Proc. SPIE,* San Jose, CA, *Vol. 3299,* pp. 274-281.

YEU98 Yeun, M., and Wu, H. R., (1998). A survey of hybrid MS / DPCM / DCT video coding distortions. *Signal processing, Vol. 70* (3), pp. 247-278.

YOU75 Young, L.R., Sheena, D., (1975). Survey of eye movement recording methods. *Behavioural Research Methods and Instrumentation*, *Vol.. 7* (5), pp. 397-429.