Stars in their Eyes: What Eye-Tracking Reveals about Multimedia Perceptual Quality

Stephen R. Gulliver, *Member, IEEE*, and Gheorghita Ghinea, *Member, IEEE*

***Abstract*— Perceptual multimedia quality is of paramount importance to the continued take-up and proliferation of multimedia applications: users will not use and pay for applications if they are perceived to be of low quality. Whilst traditionally distributed multimedia quality has been characterised by Quality of Service (QoS) parameters, these neglect the user perspective of the issue of quality. In order to redress this shortcoming, we characterise the user multimedia perspective using the Quality of Perception (QoP) metric, which encompasses not only a user's satisfaction with the quality of a multimedia presentation, but also his/her ability to analyse, synthesise and assimilate informational content of multimedia. In recognition of the fact that monitoring eye movements offers insights into visual perception, as well as the associated attention mechanisms and cognitive processes, this paper reports on the results of a study investigating the impact of differing multimedia presentation frame rates on user QoP and eye path data. Our results show that provision of higher frame rates, usually assumed to provide better multimedia presentation quality, do not significantly impact upon the median coordinate value of eye path data. Moreover, higher frame rates do not significantly increase level of participant information assimilation, although they do significantly improve overall user enjoyment and quality perception of the multimedia content being shown.**

***Index Terms*— Eye-tracking, Quality of Perception, Frame Rate, Multimedia Video.**

I. Introduction

THE effectiveness of distributed multimedia applications in entertainment, education and business is driven by the networking protocols and communications systems that deliver multimedia to the end-user. Research and development in network protocols is currently driven purely from a technical perspective, with little or no reference to the benefit to the user. However, the ultimate effectiveness of a multimedia presentation is measured by the user's multimedia experience in terms of satisfaction and information assimilation. Key to this is the issue of *quality* of the multimedia presentation. Quality, from our perspective, has two main facets: *of perception* (QoP) and *of service* (QoS). The latter characterises the technical side of computer networking and represents the performance properties that the underlying network is able to provide. The former encompasses the dual, *infotainment* (i.e. combined information and entertainment) nature of multimedia and characterises the human side of the distributed multimedia experience, that is, the user's ability to analyse, understand and synthesise the informational content of a presentation as well as user satisfaction with the presentation, and one of the main aims of our work is to uncover the relationship between QoS and QoP, for the ultimate benefit of the end-user.

User perception of multimedia has been studied extensively in the educational psychology [21] and HCI fields [5] [9] [23]. However, these studies aimed to improve user perception through the use of innovative educational, ergonomic and presentational techniques, optimistically assuming that the underlying network is able to provide the optimum QoS that yields an excellent level of quality of multimedia presentations. In an ever-growing number of cases, though, the underlying communication system will not be able to provide an optimum QoS due to two competing factors: multimedia data sizes and network bandwidth. This results in degraded network performance, manifesting itself through phenomena such as congestion, packet loss, bit errors and out-of-order arrivals. The focus of networking research has been preponderantly driven from the purely technical perspective of managing QoS parameters with little or no analysis made of the benefit to the user – the telling factor behind the widespread acceptance of multimedia technologies.

The study described in this paper builds upon previous work done by the authors [10][11], which researched the impact that varying QoS settings had on user QoP. Whilst, in previous work, QoP data were elicited indirectly during post-experiment user interviews, the use of eye tracking in this project enabled the recording of gaze information directly during experimentation, thus gaining access to data that would not otherwise be available.

Monitoring eye movements offers insights into visual perception, as well as the associated attention mechanisms and

Manuscript received June 4, 2003.

S. R. Gulliver is with the Department of Information Systems and Computing, Brunel University, Uxbridge, Middlesex, UB8 3PH, United Kingdom (e-mail: Stephen.Gulliver@brunel.ac.uk).

G. Ghinea is with the Department of Information Systems and Computing, Brunel University, Uxbridge, Middlesex, UB8 3PH, United Kingdom (CO 80305 USA (phone: +44-1895-274000x3635; fax: +44-1895-251686; e-mail: George.Ghinea@brunel.ac.uk).

cognitive processes. Interpretation of eye movement data can be based on the empirically validated assumption that when a person is performing a cognitive task, while watching a display, the location of his/her gaze corresponds to the symbol currently being processed in working memory [13] and, moreover, that the eye naturally focuses on areas that are most likely to be informative [19]. Hence, this would help provide a complete picture of QoP, as eye movements provide a wealth of detail about how people acquire and process information. Moreover, eye tracking is also attractive since it is possible for data to be collected with a fine temporal grain and subjects need little instruction or training to produce informative data [25].

The underlying premise of our research is that degradation in the network QoS results in loss of quality of the multimedia presentation. By monitoring user eye movements, we would have a more complete indication of the user perception and satisfaction with the multimedia presentation. For example, erratic eye movements, indicated by frequent saccades (rapid eye movements between regions of informative interest) would suggest a loss of focus and would be reflected in the user level of information assimilation and understanding of the multimedia presentation. Eye tracking thus serves to pinpoint the thresholds at which varying network QoS affects the user multimedia experience. Eye tracking would also help in elucidating questions that have arisen out of previous work on QoP, such as why people do not notice obvious informational cues, as reported in [10]. Thus, by comparing the eye movements of the test subjects with their QoP scores one could also extrapolate how the user's attention changes as a result of changes in QoS parameters [15].

The structure of this paper is as follows. Section 2 presents in more detail research related to Quality of Perception; which is followed in Section 3 by a discussion of eye-tracking topics, encompassing visual perception, eye movements and eye tracking techniques. Section 4 describes the empirical study undertaken as part of our research, while Section 5 presents the main results obtained. Finally, in Section 6, conclusions are drawn and avenues for future research based on our findings are proposed.

II. QUALITY OF PERCEPTION (QOP)

A. *Why Quality of Perception?*

In order to explore the human side of the multimedia experience, we have used the notion of QoP (Quality of Perception). QoP is a term which encompasses not only a user's satisfaction with the quality of multimedia presentations (denoted by QoP-S), but also his or her ability to analyse, synthesise and assimilate the informational content of multimedia (denoted by QoP-IA). QoP-S is subjective in nature and, in our work, consists of two component parts: QoP – LOE (the user's Level Of Enjoyment whilst viewing multimedia content) and QoP – LOQ (the user's jugdement concerning the objective Level Of Quality assigned to the multimedia content being visualised).

In a distributed setting, quality of digital multimedia has traditionally been measured using QoS technical parameters, such as jitter, delay, as well as loss and error rates. Although measurable, such objective parameters disregard the user's perception of what defines multimedia quality [29] [30]. To date, there has been a common assumption in the computer networking community that many quality issues will be resolved through objective solutions, such as increased bandwidth allocation [14] [32]. The majority of QoS research has therefore been systems oriented, focusing on factors such as traffic analysis, scheduling and routing.

Due to the multi-dimensional nature of multimedia, it is impossible to rely on objective factors alone when defining multimedia quality. 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. Therefore, when evaluating multimedia quality, subjective testing by viewers must be considered in combination with objective testing.

Blakowski and Steinmetz showed that synchronisation between media is generally characterised by three regions: one in which synchronisation errors are unnoticeable by the user, one in which they are perceived but tolerated, and one in which they are found irritating [2] [30]. Kawalek, on the other hand, is more interested in the cut-off rate beyond which the quality of transmitted audio and video becomes unacceptable to human users in desktop conferencing environments. He showed that the perception of media loss is highly dependent on the medium in question. While Bouch et al have researched the effect of latency on perceived Web QoS [3], Wijesekera et al build on Steinmetz's and Apteker's earlier work and investigate the perceptual tolerance to discontinuity caused by media losses and repetitions, and to that of varying degrees of mis-synchronisation across streams [33].

Apteker et al [1] studied multimedia video clips at different frame rates over a range of different bandwidths, with user preference being used to determine 'user watchability'. Apteker et al used three dimensions, which they considered inherent in all video messages: the temporal nature of the data, together with the importance of the auditory and visual components. Apteker showed that 'user watchability' was significantly affected by the content of the video clips, not just the level of available bandwidth. As 'user watchability' is defined by user preference, yet is effective over a range of different bandwidths, it acts as a limited form of QoP.

To measure the QoS impact of multimedia video clips on user perception and understanding, Ghinea and Thomas [10] presented candidates with a series of windowed (352*288 pixels) MPEG-1 video clips, between 31 and 45 seconds long. Three frame rates were used: 25, 15 and 5 fps (frames per second). The clips were chosen to cover a broad spectrum of subject matter including: spatial parameters, temporal

parameters, and importance of audio, video and textual information in context of the clip. Their results showed that:

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. In fact, in some instances the user seemed to assimilate more information.

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 value of such presentations seemed to be consistent, irrespective of the frame rate at which they are shown. The link between entertainment and understanding was found not to be direct.

Ghinea and Thomas [10] introduced the notion of QoP, as they concluded that objective factors alone were incapable of defining the perceived quality of multimedia video. QoP uses level of 'information transfer' (objective) and user 'subjective satisfaction' (subjective) to determine the perceived level of multimedia quality.

B. Measuring QoP

To understand QoP in the context of our work, it is important that the reader understands how objective and subjective QoP factors were defined and measured.

1) Measuring Information Assimilation (QoP-IA)

In our approach, QoP-IA was expressed as a percentage measure, which reflected a user's level of information assimilated from visualised multimedia content. Thus, after watching a particular multimedia clip, the user was asked a standard number of questions (ten, in our case) which examined information being conveyed in the clip just seen, and QoP-IA was calculated as being the proportion of correct answers that users gave to these questions.

For each feedback question, the source of the answer was determined as having been assimilated from one or more of the following information sources:

V : Information relating specifically to the video window, for example, pertaining to the activity of lions in a documentary clip.

A : Information which is presented in the audio stream.

T : Textual information contained in the video window, for example: information contained in a caption (for example: the newscaster's name).

All QoP-IA questions must have a definite answer. The following example: (from the Weather video clip used in our experiments) is an example of the variation of questions that may be used, as well as the information sources being tested:

1) What day of the week is the forecast for? WEDNESDAY (T)
2) What is the time? Both 6:54am and 6:55am were displayed (T) (T)
3) How is the weather in the central part of the Mediterranean, around Italy? Overcast / Rainy (V)(A)
4) How many distinctly different maps have been used in the clip? 2 (Europe / UK) (V)
5) Where, according to the forecaster, is there going to be sunshine? Cypus / Egypt (A)
6) Will driving be easy in the UK on that day? No (V) (A)
7) What colour is the forecaster's hair? Grey / White (V)
8) What's the weather on the coast of the U.K. like? Overcast / Rainy (V) (A)
9) What will the maximum visibility be in foggy areas of the U.K? 100M (A)(T)
10) What colour is the map of mainland Europe? What about the U.K. and Ireland? Green and Blue (V) Same as for mainland Europe. (V)

These questions have unambiguous answers, making it possible to determine if a participant had answered this question correctly or incorrectly. Since, in our experiments, questions can only be answered if information is assimilated from specific information sources, it is possible to determine the percentage of correctly answered questions that relate to the different information sources within specific multimedia video clip.

Thus, by calculating the percentage of correctly answered question from different information sources, it was possible to generalise from which information sources participants absorbed the most information. Using this data it is possible to determine and compare, over a range of different multimedia content, potential differences that might exist in QoP-IA.

2) Measuring Subjective Level of Enjoyment (QoP-LOE)

The subjective Level of Enjoyment (QoP-LOE) experienced by a user when watching a multimedia presentation, was polled by asking users to express, on a scale of 0 - 5, how much they enjoyed the presentation (with scores of 0 and 5 representing "no" and, respectively, "absolute" user satisfaction with the multimedia video presentation). This information was also subsequently used to determine whether ability to assimilate information has any relation to user level of enjoyment, one of the two component parts of QoP-S, the second essential constituent (beside information analysis, synthesis and assimilation) of QoP.

3) Measuring Subjective Level of Quality (QoP – LOQ)

The other component part of QoP-S is, in our approach, the subjective Level of Quality. In order to measure this, 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).

III. EYE TRACKING

A. Introducing the Human Visual System

Light reflected from objects in the visual field enter the eye

and pass through the lens, which projects an inverted image of the object onto the retina at the back of the eye. The retina consists of approximately 127 million light-sensitive cells (120 million are called rods, 7 million are called cones). Cones are less light sensitive than rods, but are responsible for capturing colour within the human visual system.

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. Movement of the eye, head and body are used to bring regions of interest 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 [28].

B. Visual Attention

The process of visual attention can be broken into two sequential stages: the pre-attentive stage and the limited-capacity stage [12][31]. In the pre-attentive stage, information is processed from the whole visual field in parallel. It is the pre-attentive stage that determines regions of interest within the visual field (defining important visual cues) and based on this pre-attentive mapping the limited-capacity stage performs high-level serial processes that are dependent on high-level search criteria. When items pass from the pre-attentive stage to the limited-capacity stage, these items are considered as selected [34].

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 lines, curvature, orientation, colour and motion [27]. The pre-attentive stage of vision operates without capacity limitations and works in parallel across the entire visual field. Learnt visual schemas therefore define how visual primitives are grouped into ‘chunks’ and how these ‘chunks’ are then perceived as objects.

2) Limited-Capacity Stage

Although, the eye naturally fixates on areas that are most likely to be informative [16], eye-gaze scan-path measurement shows that the definition of ‘informative’ is dependent on the user’s current mental process. Four distinct looking states have been defined [14]:

Spontaneous looking: when a subject is not actively looking for, or thinking about, any specific object. For example: 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. For example: when a subject thinks of a object within their visual field, (s)he 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.: eye-tracker controlled graphical user interface.

3) Three-way interaction: Eye-Tracking, QoP and QoS

When the pre-attentive and limited-capacity 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. The principal method for moving the eyes 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 saccade the processing of the visual image is suppressed, therefore processing of the retinal scene occurs mainly between saccadic periods called fixations, which last between 200-600ms. Eye-tracking equipment is accordingly categorised as being either fixation or saccade pickers, depending on their particular capture method [14].

Indeed, eye tracking is increasingly being used as a tool for obtaining information about human perceptive and cognitive processes [17][25], as it is based on the empirically-validated assumption that the eye naturally centres on areas that are most likely to be informative. Thus, Mackworth and Bruner [19] 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 [19][20]. Mackworth and Morandi noted that informative areas are identified within the first two seconds of observation [20], a conclusion that has been reported in other studies of eye movement [6][34].

Moreover, 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 [24], 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 [3], eye-typing [26], and even mouse clicking, through the development of an ‘eye-mouse’ [18]. 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 [22]. Eye tracking is also currently being used in virtual reality-based education and training, ranging from such diverse topics as aircraft inspection [7] to driving [29]. However, in the context of this paper, we are interested in the relationship between eye movement and user perception of multimedia.

The relationship between eye movement and user perception of multimedia has been investigated in [8] and [24].

The former study explores both visual attention (given by eye tracking patterns) and information recall of subjects being presented with a single multimedia educational application, displayed with optimum QoS parameters. The authors then went on to propose a series of guidelines to be used in web animation based on 'contact points' (co-references between text and animation obtained from the initial eye tracking study) [9]. Thus, the first study only focuses on the informational assimilation component of QoP, neglecting the satisfaction side of the multimedia experience. Moreover, users were shown only one particular type of multimedia category with constant, optimum QoS parameters, which fails to reflect the multitude of multimedia applications and the variety of prevailing network conditions that exist in the three-way interaction between eye-tracking, QoP and QoS. The second study investigated the effect that multi-resolution displays have on user perception. The idea behind this piece of research was to reduce the use of resources by not presenting a uniform level of visual detail across the whole display area of a screen, but rather to render a high level of visual detail only around the centre of the user's gaze. Whilst a spatial QoS parameter was indeed varied, based on user eye gaze, it was felt that this study neglects the concept of information assimilation and the multimedia diversity that QoP inherently possesses. Although both studies point to the fact that adaptive multimedia presentations, based on eye tracking results, can result in enhancement of the user multimedia experience, to the best of our knowledge no-one has examined the three-way interaction between eye tracking, QoP and QoS.

IV. Experiments

A. Participants

Our study involved 36 participants, who were evenly divided into six experimental groups: 1a, 1b, 2a, 2b, 3a and 3b. Participants were aged between 21 and 55 and were taken from a range of different nationalities and backgrounds. All participants spoke English to a degree-level qualification, were computer literate, and were presented with a series of 12 windowed MPEG-1 video clips, each between 31 and 45 seconds long.

B. Experimental Material

The multimedia video clips used in this experiment were chosen to cover a broad spectrum of infotainment subject matter. Multimedia video clips vary in nature from those that are informational in nature (such as a news / weather broadcast) to ones that are usually viewed purely for entertainment purposes (such as an action sequence, a cartoon, a music clip or a sports event, as detailed in Fig. 1). Specific clips were chosen as a mixture of the two viewing goals, such as the cooking clip).

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

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.

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.

Cooking clip - although largely static, there is a wealth of culinary information being passed on 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.

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.

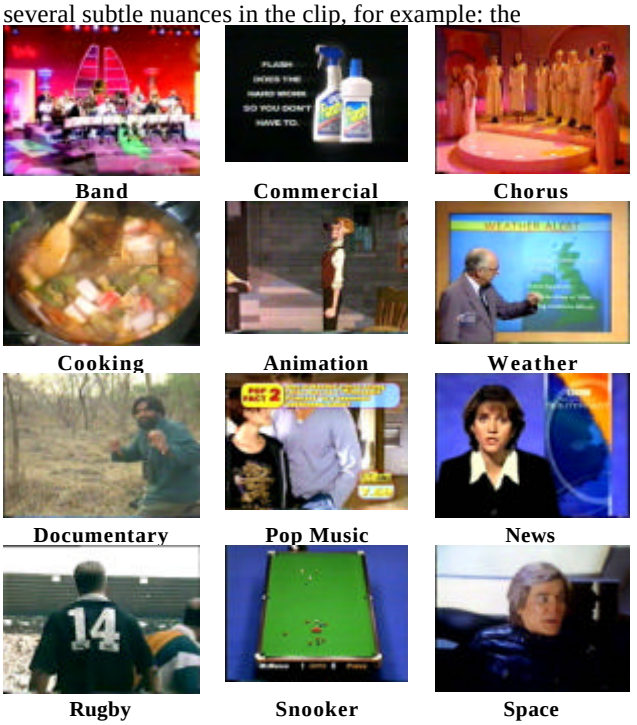

Fig. 1. Shows video frame 600, for the 12 video clips used in our experiment, demonstrating the diversity of multimedia being considered.

Weather clip - this is a clip about forthcoming weather in Europe and the U.K. 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 by the oral presentation of the forecaster.

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

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.

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.

Rugby clip - presents a test match between England and New Zealand. Essential textual information (the score) is displayed in the upper left corner of the screen. The main event captured is the score of a try. As is expected, the clip is characterised by great dynamism.

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) clearly displayed on the screen.

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).

C. Experimental Set-up

In our experiment, only one QoS parameter – frame rate – was varied. We were particularly interested in frame rate, as the frame rate with which a multimedia presentation is shown is the one parameter that has the greatest bandwidth implications in today's distributed multimedia systems – and bandwidth is arguably the most scarce networking resource in such environments. Accordingly, a within-subjects design was chosen, where participants viewed each clip of our study 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. Moreover, in order to counteract any possible order effects, the video clips were shown in a number of order and frame-rate combinations dependent on the defined experimental group name (e.g. group 1a, group 2b, etc.). For example, as detailed in Table 1, participants in group 1b were shown video frame-rates, as defined under the '**Group 1**' heading, yet the videos were shown in B first order: Documentary through Space (**B**), then Band through Weather (**A**). As can be observed, across the six experimental groups all possible combinations of frame-rate were shown in both A and B orders.

TABLE 1: FRAME RATE AND VIDEO ORDER PRESENTED TO EXPERIMENTAL GROUPS.

	Video	Group 1	Group 2	Group 3
A	Band (Jazz Band)	25	5	15
A	Commercial	5	15	25
A	Chorus - Choir	15	5	25
A	Cooking	15	25	5
A	Animation	25	15	5
A	Weather	5	25	15
B	Documentary	5	15	25
B	Pop	15	25	5
B	News	5	25	15
B	Rugby	25	5	15
B	Snooker	15	5	25
B	Space	25	15	5

To ensure that experimental conditions remained constant throughout, consistent environmental conditions were used for all participants. An Arrington Research, Power Mac G3 (9.2) infrared camera-based pupil tracking, ViewPoint EyeTracker was used in combination with QuickClamp Hardware (Fig. 2). 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 – Fig. 2). The position of the chin rest and camera were, however, changed dependent on the specific facial features of the participant. For further technical information concerning the set-up of the ViewPoint Eye-tracker see Table 2. To avoid audio and visual distraction, a dedicated uncluttered room was used throughout all experiments. 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. State transition scripts were developed and implemented in the ViewPoint software.

a) **b)**

Fig. 2. a) Power Mac G3 (9.2) ViewPoint EyeTracker, used in combination with b) QuickClamp Hardware.

TABLE 2: TECHNICAL SPECIFICATION OF VIEWPOINT EYE-TRACKER..

Accuracy	Approximately 0.5° - 1.0° visual arc
Temporal 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	Automatic blink detection and suppression.
Data recorded	Eye data: X, Y position of gaze, delta time, and regions of interest. Asynchronous data recorded include: State transition markers and key presses, data from other programs. Data is stored in ASCII files.

Transition scripts allow movement through a number of

defined states and are dependent on participant key presses. They allow each experiment to proceed at a flexible rate, marking relevant experimental points, such as the start of the video or a key-press, on the stored eye-tracking data

D. Experimental Process

Each participant was 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 eye-tracker screen without spectacles. Participants wearing contact lenses were not asked to remove lenses, however, special note was made and extra time was given when mapping the surface of the participant's eye to ensure that pupil fix was maintained throughout the entire visual field. After each participant was given a brief introduction, the ViewPoint system was loaded and the participant was asked to place their nose in the QuickClamp nose-rest and their forehead on the forehead rest, thus removing risk of rotation or tilt during the study session. As the shape and colour shades of participants' facial features varied considerably, time was taken to adjust the chin-rest, infrared red capture camera and software settings to ensure that pupil fix was maintained in the 'Eye Camera Window' throughout the entire visual field (see Fig. 3). Once configuration set-up was complete, automatic calibration was made using a full screen stimulus window. However, point re-calibration was also used if an error, such as head movement, caused a non-smooth pupil mapping in eye-space window (see Fig. 3).

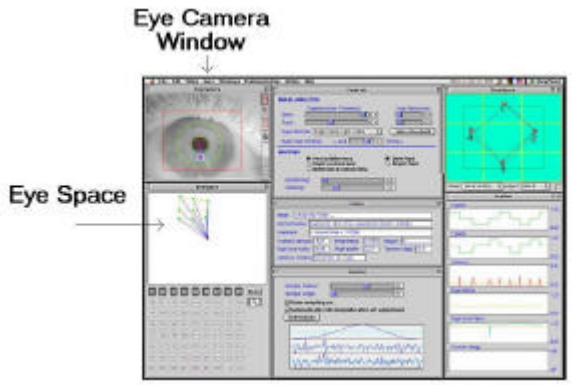

Fig. 3. Layout of ViewPoint software - developed by Arrington Research Inc.

When calibration was complete, eye-space settings were stored and a new data file was created for eye-tracking data. The stimulus window was expanded and the relevant state transition script was loaded and activated. Before viewing each multimedia video the participant was asked to get into a comfortable position and place his/her chin on the chin-rest. By focusing on a temporary spot in the centre of the screen, errors, caused due to slight shifts in head position, were corrected. The participant was reminded to keep his/her head still for the duration of the video. The presentation state was incremented, using a key-press command (+), which both added a transition marker to the eye-tracking data and started the next video clip.

After showing each video clip, the video window was closed and the participant was asked QoP questions about the video that they had just seen. QoP questions were chosen to encompass both objective (QoP-IA) and subjective (QoP-LOE and QoP-LOQ) aspects of the information presented in the specific clip. The questions were designed to examine the type of information assimilated by the user in accordance with the QoP definition.

E. Extracting Frame Based Eye-Tracking Coordinates

Data samples contained: x values, y values and timing data, making data extraction three dimensional in nature. Using the delta time between eye-tracking samples, we were able to calculate the relative participant eye-position for specific video frames. After manipulation, data extracted from each eye-tracking data sample included: frame number, delta time (expressed in ms), x coordinate (range: 0-10000), y coordinate (range: 0-10000) - see Fig. 4. X and y coordinate values (range 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 comparison between participant sets, eye-tracking data was sampled at 25Hz for all clips used as part of our experiments, corresponding to the maximum frame rate being displayed.

3	:Movies:ModelFest:BA05.mpg		
0	0	4896	4896
1	40	4896	4896
2	80	4896	4896
3	120	4896	4896
4	160	4896	4896
5	200	4896	4896
6	240	4896	4896
7	280	4896	4896
8	320	4896	4896
9	360	4896	4896
10	400	4896	4896
11	440	4896	4896
12	480	4896	4896

Fig. 4. Frame-based eye-tracking data - Band Video Clip (5fps).

Frame based eye-tracking coordinates for all participants were saved in separate files, thus representing the eye-tracking data for a specific participant viewing a specific video clip (432 files in total). For each of the 12 videos, at the three defined frame-rates (5, 15, 25 fps), 12 sets of eye-tracking data were recorded (see Table 1). Data relating to each of the videos, at all of the frame-rates, were combined so that x and y coordinates for participant relating video frames could be analysed (Fig. 5).

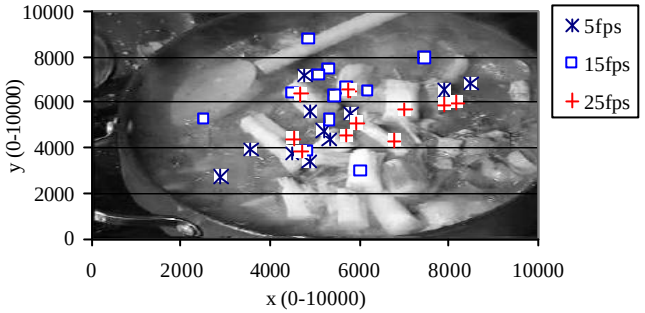

Fig. 5. Scatter graph for participants (12 – 5fps, 12 – 15fps, 12 – 25fps), viewing cooking clip - frame number 95.

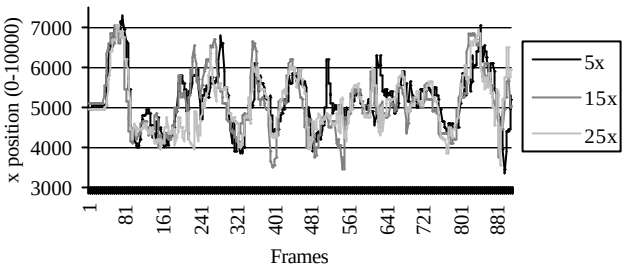

Figure 6: Space Action Movie x-coordinate video eye-path.

V. RESULTS

A. *Impact of Frame-rate on Video Eye-Paths*

To allow us to statistically correlate eye-position between frame rates (5, 15 and 25 fps) over the duration of the video clip (between 650-1000 frames), three coordinate points were required for each frame, each one relating to a specific frame-rate group (5, 15 and 25 fps). As we are not aware of any previous eye-tracking data analysis across multiple frames, no precedent for summarising participant group eye-tracking data was discovered by the authors. Therefore, to avoid inclusion of extreme outlying points whilst removing unwanted data, such as error coordinates as a result of participant blinking, we determined that the coordinate points for each video frame should be the median value of the data within each of the participant groups (5, 15 and 25 fps). Although a median value is not ideal, especially if multiple regions of interest occur within a frame, the authors considered it to be least prone to error values. 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 (video eye-path), for clips shown at each of the available frame-rates (Fig. 6). The example used in Fig. 6 shows the x coordinate value for the band video clip. The band clip shows a dynamically changing music video for a jazz band (Fig. 1). Although eye fixations tend to return to 5000, 5000 (the centre of the screen) in this example, this is not always the case. We can therefore assume that this trend is clip dependent. Mapped values represent two of the three data dimensions (one coordinate value and time) and therefore, whilst allowing analysis, reduce statistical complexity.

Statistical correlations (Kendall's tau-b and Spearmans 2–tailed nonparametric tests) were carried out between median coordinate values, for frame rates 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), for all of the 12 multimedia video clips (72 tests in total). This allows us to determine whether x and y coordinate values, from specific video clips shown at varied frame rates, statistically correlated, i.e. similar median trends of eye movement occur for groups of people shown the same video clip at different frame rates. All 72 correlation tests showed a significance correlation value of $p<0.001$ between the video eye-paths across the different frame rates, which implies a strong correlation between the median eye-position of participants, independent of the video being presented.

This result shows that, for median coordinate values mapped across time, eye movement significantly correlates independent of the video frame rate. With such strong correlation between participants, and the fact that strong correlation exists for each of the diverse multimedia video clips, we can conclude that frame rate does not significantly impact median video eye-path. Although by reducing the frame rate we reduce the level of information within the peripheral visual area, irrespective of multimedia content, our results suggest that overall median eye-path is consistent independent of the presentation frame rate. Significantly correlating video median eye-paths suggest that a eye path exists that is not affected by frame rate variation. If a video specific eye-path exists, then it may be possible to use this in combination with adapted multimedia to improve user perception of the video presentation. Previous work has shown that video enhancement around the viewed area can cause an improvement in user perception of video [24]. Therefore, if video specific median eye-path trends exist, manipulation of video, independent of frame rate, could be used to enhance areas around the median coordinate values.

B. *Impact of Frame-rate on user QoP*

1) *Objective QoP: QoP-IA*

Questions used to measure QoP-IA are specific to the video clip being viewed. If the same questions are used, when similar participants groups are shown the video clips at different frame rates (5, 15 and 25 fps), any significant difference between

QoP-IA would suggest a perceptual change as a result of frame rate variation.

Papers examining the QoP-IA were marked, allowing the type and source of information assimilated by each participant to be determined. This allowed the subsequent calculation of average levels of video (V), audio (A), and textual (T) information assimilated by the participants in each (*frame_rate, video_clip*) category, as is detailed in Table 3.

To statistically measure whether there is any significant difference, caused by a change in frame rate, an Analysis of Variance (ANOVA) test was carried out with frame rate as an independent variable and V,A,T, as dependent variables.

This analysis showed that assimilation of video, audio and textual information was not significantly affected by the frame rate variation. This suggests that the level and type of information assimilated by participants was not significantly affected by a change in frame rate, a finding that supports previous work [10] (Fig. 7). We can conclude that QoP-IA is not significantly impacted by a change in video frame rate. This result facilitates the manipulation of presentation frame rate as a means of bandwidth reduction, without impacting participant QoP-IA.

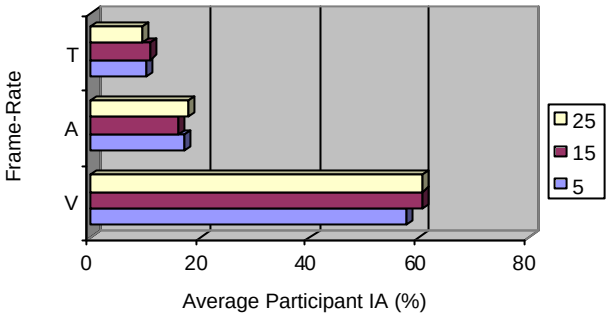

Figure 7: QoP-IA breakdown showing average participant information assimilation from (V)ideo, (A)udio and (T)extual sources,when viewing multimedia clips at 5, 15 and 25fps.

2) Subjective QoP: QoP-LOE, QoP-LOQ

ANOVA tests were also carried out to statistically measure whether there is any significant difference, caused by a change in frame rate, in LOE and LOQ. These highlighted that both QoP-LOE {$F(1,2) = 4.482$, $p=0.012$} and QoP-LOQ {$F(1,2) = 6.911$, $p=0.001$} were significantly different when shown at different frame rates (Fig. 8), confirming the results of Apteker et al [1] (who, however, only looked at QoP-LOE). Post-hoc Tukey tests showed, however, that the differences in QoP-LOE and QoP-LOQ were not statistically significant between the cases of multimedia video content being viewed at 15 and 25 fps.

TABLE 3: AVERAGE QOP-IA SCORE FOR ALL PARTICIPANT GROUPS (VIDEO, AUDIO AND TEXTUAL). N/A = NON-APPLICABLE INFORMATION AVAILABLE.

	Video IA (%)			Audio IA (%)			Textual IA (%)		
	5	**15**	**25**	**5**	**15**	**25**	**5**	**15**	**25**
Band	56.66	58.3	55.8	33	41	58	N/A	N/A	N/A
Commercial	73.71	66.57	75	30	22	30.33	N/A	N/A	N/A
Chorus	64.77	66.66	67.55	N/A	N/A	N/A	33	50	41
Cooking	59.1	54.1	54.1	33.28	33.28	32.14	N/A	N/A	N/A
Animation	57.33	60.11	67.55	33.28	46.42	50	N/A	N/A	N/A
Weather	78.2	90	88.2	46.6	51.6	48.2	45.75	39.5	52
Documentary	69.1	75.8	66.6	50	33	33	N/A	N/A	N/A
Pop Video	50	70.14	66.57	47	58.33	38.66	35.25	33.25	29
News	62.5	50	66.5	63.88	48.11	51.77	70.5	62.5	66.5
Rugby	47.5	43.3	50	0.4	0	3.3	82.5	81.25	60.25
Snooker	55.12	66.6	59.37	66	41	83	66.5	68.75	56.25
Space	52.5	57.5	53.3	25	10	66	N/A	N/A	N/A

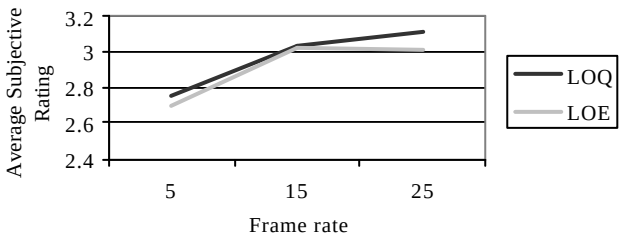

Figure 8: Average QoP-LOE (enjoyment) and QoP-LOQ (Quality), when viewed at 5, 15 and 25 fps.

Although median video eye-path and level of information assimilation (QoP-IA) are not significantly affected by varied frame rate, the result shows that a user's perception of quality and enjoyment is significantly affected by a change in frame rate and shows that users are subjectively aware that a change in frame rate occurs. It is interesting, however, that a change in frame rate, which causes a significant difference in subjective QoP-variables, does not cause a change in the objective level of information assimilation (QoP-IA). This implies that manipulation of presentation frame rate, as a means of bandwidth reduction, although not significantly impacting participant QoP-IA, will significantly affect a participant's level of QoP-S (Subjective = Quality and Enjoyment).

A. Impact of Clip Type on user QoP

1) Objective QoP: QoP-IA

QoP by its nature is video specific and is determined using questions that accept the fact that there are differences in the information distributed by the possible range of multimedia content visualized. As expected, ANOVA with clip type as an independent variable showed that level of QoP-IA (information assimilation) for video {$F(1,11) = 39.533$, $p<0.001$}, audio {$F(1,11)= 79.724$, $p<0.001$} and textual {$F(1,11) = 82.193$, $p<0.001$} sources was significantly different for different video clips (Fig. 9). *Video IA:* On average participants successfully assimilated the greatest level of information from the video information. It is interesting that QoP-IA remains largely consistent, independent of frame-rate. *Audio IA:* The level of information assimilated from audio information sources varies considerably and is dependent on the video clip. It is interesting to note that participants were consistently unable to recall information from the rugby audio, yet were able recall a high level of video and textual information. *Textual IA:* A number of video clips do not have textual information, however, the level of textual information assimilated from multimedia presentation appears to vary with the clip.

2) Subjective QoP: QoP-LOE, QoP-LOQ

ANOVA with clip type as an independent variable highlighted that both QoP-LOE {$F(1,11) = 3.095$, $p<0.001$} and QoP-LOQ {$F(1,11) = 5.425$, $p=0.001$} were significantly different for different video clips. This shows that user perception of what is enjoyable and what is 'quality' is significantly affected by the video clip being viewed.

The level of enjoyment varies considerably depending on both the video clip type being viewed and the frame-rate at which the video is being shown (Fig. 10). For example: The animation video clip appears to be consistently entertaining, yet a small reduction can be seen when showed at five frames per second. The rugby clip displays the largest variation in level of enjoyment, as a result of varied frame-rate, ranging from 1.83 to 3.12 (out of 5) for 5 and 25 fps respectively.

The variation in level of quality is an interesting result, as significantly different scores occur as a result of purely video clip-type. If a user's subjective perception of 'objective video quality' is dependent on the video clip being shown, it questions the usage of purely objective parameters alone when defining video quality.

VI. Conclusions

In this paper, the three-way interaction between perceptual multimedia quality, eye-gaze location and frame rates has been explored. Recognising the infotainment duality of multimedia

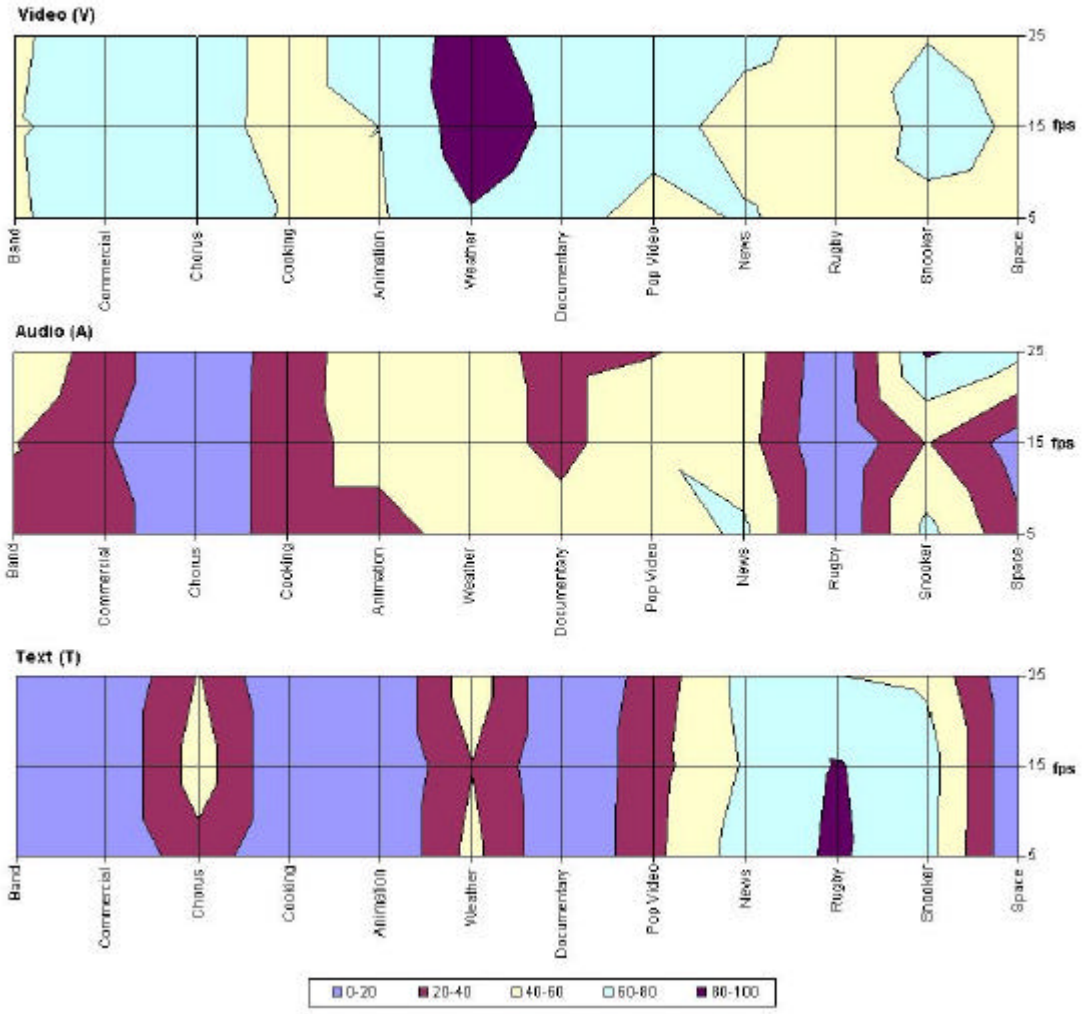

Figure 9: Average percentage QoP-IA scores for all information sources, for all frame-rates.

presentations, perceptual multimedia quality was evaluated using the Quality of Perception – QoP - measure, which encompasses both the user's ability to analyse, understand and synthesise the informational content of a presentation as well as user subjective satisfaction with the presentation. We have thus used both QoP and median eye-tracking data, across multiple frame rates (5, 15 and 25 fps), to analyse the impact that varied frame rate (and, implicitly, bandwidth) has on user perception and video eye-path.

Our results highlighted that, independently of subject matter, frame rate does not significantly impact the median video eye-path. This suggests that the location of a user's focus of attention does not significantly change if (s)he is presented with what is technically recognised as a better quality multimedia presentation (i.e. a multimedia presentation with a higher frame rate), and questions the premise that bandwidth-constrained environments (such as wireless communications) are fundamentally incapable of supporting (perceptually) good quality multimedia applications.

When further investigating this result by considering the associated QoP data, we observed that, whilst the information assimilation component of QoP was unaffected by different frame rates, the subjective satisfaction with the multimedia quality of the clips considered in our experiments was indeed affected by the frame rate of the presentation. If video specific median eye-paths exist, independent of frame rate, then participants without perceptual limitations [11] will view similar regions of interest and therefore objectively assimilate similar levels of information. This finding opens new avenues for further research, for it implies that having region-of-interest based presentations of multimedia content, where perceptually relevant regions are played at higher frame rates than the surrounding areas, could potentially improve overall QoP.

Moreover, our results support the view that definition of 'multimedia quality' cannot be defined purely using QoS parameters, for our work has provided evidence that different participants enjoy and perceive different things to be of 'quality'. Whilst these findings may be unsurprising they do demonstrate that subjective factors should be considered more relevant when defining or evaluating 'video quality'.

Limitations in our experimental set-up meant that dwell time

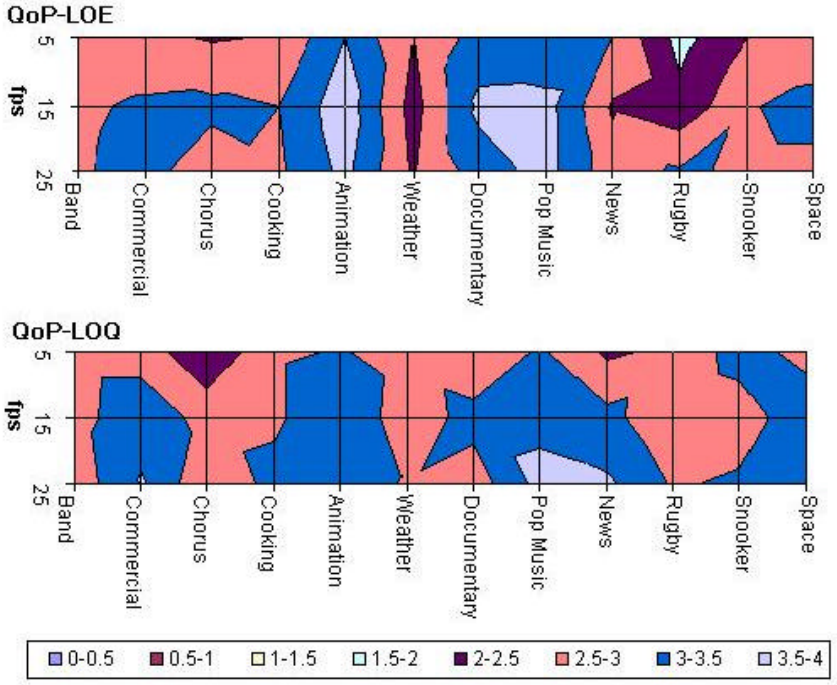

Figure 10: Average QoP-S factors for all clips, across all frame-rates.

was not included as a variable on the output data files and accordingly is not discussed in this study. Accordingly, the authors see a specific need for further research to analyse the impact of frame rate variation and clip type on fixation dwell time.

In concluding, we identify that multimedia quality definition needs to be reconsidered. The suggestion of video specific eye-paths needs to be further investigated and the impact of multimedia adaptation, around the video eye path, needs to be measured. The fact that frame rate significantly impacts a user's subjective definition of quality and enjoyment, yet has no significant effect on the information assimilation component of QoP, has implications on using purely objective testing when defining multimedia quality. Further work is required to identify the impact that both objective and subjective parameters have on user QoP, if future systems involving multimedia adaptation are not going to disregard a user's own definition of multimedia quality.

REFERENCES

[1] R. T. Apteker, J. A. Fisher, J. A. Kisimov, and H. Neishlos, "Video Acceptability and Frame Rate," *IEEE Multimedia*, vol. 2, no. 3, pp. 32-40, 1995.

[2] G. Blakowski, and R. Steinmetz, "A Media Synchronisation Survey: Reference Model, Specification, and Case Studies," *IEEE Journal on Selected Areas in Communications*, vol. 14, no. 1, pp. 5-35, Jan. 1996.

[3] A. Bouch, A. Kuchinsky, and N. Bhatti, "Quality is in the eye of the beholder", *Proceedings of the CHI 2000 Conference on Human Factors in Computing Systems*, pp. 297-304, The Hague, The Netherlands, 2000.

[4] M. D. Byrne, J. R. Anderson, S. Douglass, and M. Matessa, "Eye Tracking the Visual Search of Click-Down Menus", in *Proc. of ACM CHI '99*, Pittsburgh, Pennsylvania, USA, 1999, pp. 402-409.

[5] S. K. Card, T. P. Moran, and A. Newell, *The Psychology of Human-Computer Interaction.* Hillsdale, NJ: Lawrence Erlbaum Associates, 1983.

[6] A. D. De Groot, "Perception and memory versus thought: Some old ideas and recent findings", in *Problem solving: Research, method, and theory*, B. Klinmuntz, Ed. New York: John Wiley, 1966.

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

[8] P. Faraday and A. Sutcliffe, "An Empirical Study of Attending and Comprehending Multimedia Presentations", in *Proc. of ACM Multimedia '96*, Boston, Massachusetts, USA, 1999, pp. 265-275.

[9] P. Faraday and A. Sutcliffe, "Authoring Animated Web Pages using Contact Points", in *Proc. of ACM CHI '99*, Pittsburgh, Pennsylvania, USA, 1999, pp. 458-465.

[10] G. Ghinea and J. P. Thomas, "QoS Impact on User Perception and Understanding of multimedia Video Clips", in *Proc. of ACM Multimedia '98*, Bristol, UK, 1998, pp. 49-54.

[11] S. R. Gulliver and G. Ghinea, "Impact of Captions on Deaf and Hearing Perception of Multimedia", in *Proc. of ICME '02*, Lausanne, Switzerland, 2002, pp. 970-974.

[12] J. E. Hoffman, "Search through a sequentially presented visual display", *Perception and Phychophysics*, vol. 23, pp. 1-11, 1978.

[13] M. A. Just and P. A. Carpenter, "Eye Fixations and Cognitive Processes", *Cognitive Psychology*, vol. 8, pp. 441-480, 1976.

[14] D. Kahneman, *Attention and Effort*. Englewood Cliffs, NJ: Prentice-Hall, 1973.

[15] S. Karn, "Saccade pickers" vs. "fixation pickers": the effect of eye tracking instrumentation on research, in *Proc. of the Eye Tracking*

Research and Applications Symposium, Palm Beach Gardens, Florida, USA, 2000, pp.87-88.
[16] L. Kaufman and W. Richards, "Spontaneous fixation tendencies for visual forms", *Perception and Psychophysics*, vol. 5, pp. 85-88, 1969.
[17] E. Kowler, "The Role of Visual and Cognitive Processes in the Control of Eye Movements", in *Eye movements and their role in visual and cognitive processes*, E. Kowler, Ed. Amsterdam: Elsevier, 1990, pp. 1-70.
[18] C. Lankford, "Effective Eye-gaze Input into Windows", in *Proc. of the Eye Tracking Research and Applications Symposium*, Palm Beach Gardens, Florida, USA, 2000, pp. 23-27.
[19] J. F. Mackworth, and J. S. Bruner, "How adults and children search and recognize pictures", *Human Development*, vol. 13, pp. 149-177, 1970.
[20] J. F. Mackworth and A. J. Morandi, "The gaze selects informative details within pictures", *Perception and Psychophysics*, vol. 2, pp. 547-552, 1967.
[21] R. E. Mayer, "Multimedia Learning: Are We Asking the Right Questions?", *Educational Psychologist*, vol. 32, no. 1, pp. 1-19, 1997.
[22] J. Nielsen (2001, May 20). *Eye Tracking Study of Web Readers*, [Online]. Available: http://www.useit.com/alertbox/20000514.html.
[23] D. A. Norman, "Cognitive Engineering", in *User-Centered System Design*, D. A. Norman and S. W. Draper, Eds. Hillsdale, NJ: Lawrence Erlbaum, 1986, pp. 31-61.
[24] D. Parkhurst, E. Culurciello, and E. Niebur, "Evaluating Variable Resolution Displays with Visual Search: Task Performance and Eye Movements", in *Proc. of the Eye Tracking Research and Applications Symposium*, Palm Beach Gardens, Florida, USA, 2000, pp. 105-109.
[25] J. B. Pelz, R. Canosa, and J. Babcock, "Extended Tasks Elicit Complex Eye Movement Patterns", in *Proc. of the Eye Tracking Research and Applications Symposium*, Palm Beach Gardens, Florida, USA, 2000, pp. 37-43.
[26] D. D. Salvucci, "Inferring Intent in Eye-based Interfaces: Tracing Eye Movements with Process Models", in *Proc. of ACM CHI '99*, Pittsburgh, Pennsylvania, USA, 1999, pp. 254-261.
[27] P. Salapatel and W. Kessen, "Visual scanning of triangles in the human newborn", *Journal of Experimental Child Psychology*, vol. 3, no 2, pp. 155- 167, 1966.
[28] A. F. Sanders, "Some aspects of the selective process in the functional visual field", *Ergonomics*, vol. 13, pp. 101-117, 1970.
[29] C. T. Scialfa, L. McPhee, and G. Ho, "The Effects of a Simulated Cellular Phone Conversation on Search for Traffic Signs in an Elderly Sample", in *Proc. of the Eye Tracking Research and Applications Symposium*, Palm Beach Gardens, Florida, USA, 2000, pp. 45-50.
[30] R. Steinmetz and K. Nahrstedt, *Multimedia: Computing, Communications and Applications*. Englewood Cliffs, NJ: Prentice-Hall, 1995.
[31] A. Tresman, "Features and objects in visual processing", *Scientific American*, vol. 255, no. 5, pp. 106-115, 1986.
[32] A. Watson and M. A. Sasse, "Multimedia conferencing via multicasting: determining the quality of service required by the end user", in *Proc. of AVSPN '97*, Aberdeen, UK, 1997, pp. 189-194.
[33] D. Wijesekera, J. Srivastava, A. Nerode, and M. Foresti, "Experimental Evaluation of Loss Perception in Continuous Media", *Multimedia Systems*, vol. 7, no. 6, pp. 486 - 499, 1999.
[34] A. L. Yarbus, *Eye movement and vision* (trans. B. Haigh). New York: Plenum Press, 1967.
[35] A. J. Glenstrup and T. Engell-Nieson, (2003, June 04). *Eye Controlled Media: Present and Future State* [Online]. Available: http://www.diku.dk/users/panic/eyegaze/article.html

Stephen R. Gulliver (M'02) received the B.Eng. (Hons) degree in microelectronics, in 1999, and M.Sc. degree in distributed information systems, in 2001, from Brunel University, United Kingdom.

He is a Ph.D. Researcher in the Department of Information Systems and Computing at Brunel University. His research interests include perceptual aspects of multimedia, accessibility, Quality of Service, as well as eye-tracking and attention analysis.

Gheorghita Ghinea (M'02) received the B.Sc. and B.Sc.(Hons) degrees in computer science and mathematics, in 1993 and 1994, respectively, and the M.Sc. degree in computer science, in 1996, from the University of the Witwatersrand, Johannesburg, South Africa; he then received the Ph.D. degree in Compuer Science from the University of Reading, United Kingdom, in 2000.

He is a Lecturer in the Department of Information Systems and Computing at Brunel University. His research interests span perpetual aspects of multimedia, Quality of Service and multimedia resource allocation, as well as computer networking and security issues.