The Perceptual and Attentive Impact of Delay and Jitter in Multimedia Delivery

Stephen R. Gulliver and Gheorghita Ghinea

***Abstract*—In this paper we present the results of a study that examines the user's perception—understood as both information assimilation and subjective satisfaction—of multimedia quality, when impacted by varying network-level parameters (delay and jitter). In addition, we integrate eye-tracking assessment to provide a more complete understanding of user perception of multimedia quality. Results show that delay and jitter significantly affect user satisfaction; variation in video eye path when either no single/obvious point of focus exists or when the point of attention changes dramatically. Lastly, results showed that content variation significantly affected user satisfaction, as well as user information assimilation.**

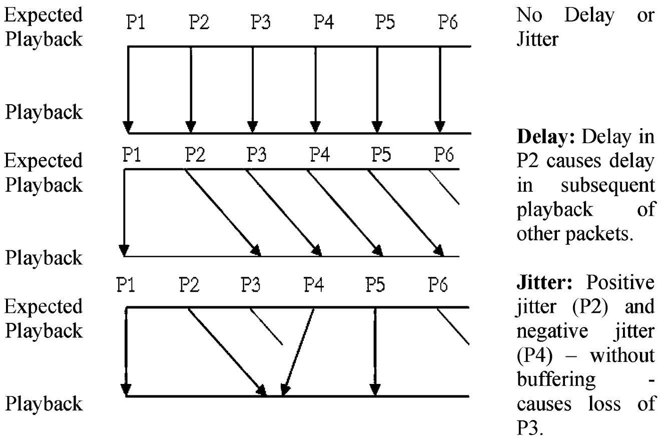
***Index Terms*—Computer interface human factors, multimedia communication, user centered design.**

I. Introduction

Delay, JITTER, and loss are important factors in the context of real-time distributed multimedia communications [1]. Whilst there is an abundance of research work investigating ways of managing these Quality of Service (QoS) parameters [2]–[4], their perceptual impact on the user has, with the exception of loss [5], [6] been largely ignored. With the emergence and proliferation of ubiquitous multimedia and interactive, content-rich, broadcast applications, it is our opinion, that, as users are 'consumers' of distributed multimedia applications and ultimately determine such applications' take-up and success, the user quality perspective is an issue that should not be ignored. Accordingly, in this paper we focus on precisely this issue and investigate the perceptual impact of delay and jitter degradations on the user.

Delay is the time taken by a packet to travel from the sender to the recipient [4]. A delay is always incurred when sending distributed video packets; however the delay of consecutive packets is rarely constant, with variation in delay defining jitter (See Fig. 1).

Although delay and jitter can be reduced via complex QoS management and/or buffering techniques, they can not be removed completely in broadcast environments. Both delay and jitter are closely linked to synchronization, which in the context of broadcast multimedia comprises the temporal relationships among media types. In a multimedia context this definition can be extended such that synchronization comprises content, spatial and temporal relations between media objects. Unfortunately, network-level errors disrupt such relationships, affecting user perception of multimedia presentations. Numerous studies concerning the perceptual impact of delay and jitter have been made. In summary, these studies show that:



No Delay or Jitter

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

Jitter: Positive jitter (P2) and negative jitter (P4) – without buffering - causes loss of P3.

Fig. 1. The effect of delay and jitter on video playback.

- Jitter degrades video quality as much as packet loss [1].
- 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 [6].
- Perceived quality of low temporal aspect video is not impacted in the presence of jitter as much as video high temporal aspect [1], [6]–[8].
- There is a strong correlation between the average number of quality degradation events (points on the computer 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 affected by jitter and packet loss [1], [6], [9]–[11].
- Momentary rate variations in the audio stream, although initially amusing, are soon deemed to be annoying. This results in participants concentrating more on the audio defect, rather than the audio content [1].

In addition, the presence of video delay can impact the synchronization of audio and video streams, and Steinmetz [8] has done important work identifying the minimal synchronization errors that have been found as perceptually acceptable. Multimedia, however, is produced for purposes that are both informative and entertaining. Accordingly multimedia user-perception must consider this *infotainment* duality. Although the impact

Manuscript received April 19, 2006; revised December 20, 2006.

The authors are with the School of Information Systems, Computing and Mathematics, Brunel University, Uxbridge, UB8 3PH, U.K. (e-mail: Stephen.Gulliver@brunel.ac.uk; George.Ghinea@brunel.ac.uk).

Color versions of one or more of the figures in this paper are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/TBC.2007.896955

of delay and jitter has been considered by other authors, these studies fail to assess the infotainment duality of the user-perspective and therefore fail to measure both the level of user information assimilation and the user's satisfaction. In this paper we focus our attention on both aspects of the user experience, employing eye tracking as an additional means of monitoring user interaction with distributed multimedia applications.

The structure of this paper is as follows: we start in Section II, by introducing the Quality of Perception concept, which is used in our experiments to assess the user's perception of quality. In Section III we introduce and justify the use of eye-tracking technology in our experiments, while the experiments themselves are described in Section IV. Results are presented and discussed in Section V, with conclusions and areas for further work being identified in Section VI.

II. Quality of Perception: An Adaptable Approach

A. *Defining Quality of Perception*

Distributed multimedia applications are produced for the enjoyment and/or education of human viewers, so the user-perspective is important to any quality definition. Previous research [2], [3], [8], [13] shows that when defining user satisfaction it is important to consider the infotainment duality of multimedia: the ability to transfer information to the user, yet also provide a level of entertainment satisfaction.

In order to explore the human side of the multimedia experience, we have used the *Quality of Perception* (QoP) concept [8]. QoP is a concept that captures the multimedia infotainment duality and more closely reflects multimedia's infotainment characteristics, i.e. that multimedia applications are located on the informational-entertainment spectrum. QoP is based on the conclusion that technical measurement alone is incapable of defining the perceived quality of multimedia video, especially when incorporating user satisfaction [2], [8], [13], [14].

QoP (as defined in Section II-B) uses level of 'Information Assimilation' (QoP-IA) and user 'satisfaction' (QoP-S) to determine the perceived level of multimedia quality. To this end, QoP is a term used in our work that encompasses not only a user's satisfaction with the quality of multimedia presentations (QoP-S), but also his/her understanding, that is the ability to analyse, synthesize and assimilate the informational content of multimedia (QoP-IA).

B. *Measuring QoP*

To understand QoP in the context of our work, it is important that the reader understands how QoP factors are defined and assessed. These issues shall now be addressed.

Measuring Information Assimilation (QoP-IA): QoP-IA implements *content query* (where test subjects are asked questions about the content of video clips after watching them) and allows us to assess a user's ability to understand/assimilate the content of multimedia video content. QoP-IA is expressed as a percentage measure that reflects the user's level of information assimilation when viewing multimedia content. Thus, after watching a particular multimedia clip, the user is asked a number of questions that examine the information being assimilated from certain information sources.

QoP-IA questions are designed so that specific information must be assimilated and understood in order to correctly answer each question; moreover, questions have unambiguous answers. QoP-IA is then expressed as a percentage measure reflecting the proportion of correct answers given by a participant (out of the total number of possible correct answers). In our study QoP-S is subjective in nature and consists of two component parts, measured using two *quality opinion scores* (where test subjects are asked for an opinion score after watching a video clip): QoP-LoQ (Level of Quality: relating to the user's subjective judgment concerning the media's objective Level of Quality) and QoP-LoE (Level of Enjoyment: relating to the user's Level of Enjoyment when viewing specific multimedia content).

Measuring Subjective Level of Quality (QoP-LoQ): In order to assess QoP-LoQ (the user's subjective judgment concerning the media's objective Level of Quality), 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).

Measuring Subjective Level of Enjoyment (QoP-LoE): To assess QoP-LoE (the user's Level of Enjoyment when viewing specific multimedia content), 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). A 6-point scale was chosen for QoP-LoQ and QoP-LoE to purposefully prevent participants expressing neutral opinions.

III. Eye Tracking

Quality is in the eye of the beholder—in our work we have taken this saying literally and have used eye tracking to monitor user gaze patterns and thus obtain a more complete characterization of user perceptual quality.

A. *Why Eye Tracking?*

The use of eye tracking is motivated by the $WYS <> WYG$ (What You See is not What You Get) relationship [5], which implies that users sometimes appear not to notice obvious informational cues in multimedia video content. Instead, in these cases users sometimes appear to determine their conclusions as a result of reasoning, arriving at them based on intuition, past experience or pre-knowledge. Consequently the use of eye tracking was proposed in future user-perspective studies in order to identify perceptually relevant areas of user eye gaze and thus provide an answer as to why people do not notice cues in the multimedia video material, thus providing a better understanding of the role that the human element plays in the reception, analysis and synthesis of multimedia data [15], [16]. Moreover, monitoring eye movements offers insights into user perception, as well as the associated attention mechanisms and cognitive processes, since the eye naturally selects areas that are most informative [17] in the context of high-level processes (see Table I).

In our work, it was felt that, whilst highlighting variations in user perception as a result of changes in quality parameters,

TABLE I
HIGH-LEVEL FACTORS CONTRIBUTING TO EYE-GAZE PATTERNS

High Level Factor	Description	Supporting Work
Location	25% of viewers concentrate on or around the centre of the screen.	[18]
Foreground /Background	Foreground objects are considered more contextual relevant than background objects.	[19][20]
People	Eyes, faces, mouths and hands are areas of significant importance to human gaze.	[20][21][22]
Context	Eye movements can drastically change depending on the instruction given whilst watching an image.	[20][21]

TABLE II
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.

the use of experimental questionnaires alone does not allow the continuous monitoring of user focus, which is important in understanding the cognitive state of the user. Accordingly, questionnaires fail to conclusively highlight the points where information assimilation occurs, or whether the information was assimilated from the presentation at all. A participant answering a QoP-IA question correctly suggests that the user either assimilated relevant information from a multimedia video presentation, or possessed pre-knowledge regarding the content of the video being shown. Consequently, questionnaire data alone does not provide a conclusive result.

B. *Implementing Eye Tracking*

Variations in 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 [23], [24]. Depending on the equipment, eye-tracking devices can be considered as either intrusive or non-intrusive in nature [25] and can be developed as either pervasive [26] or standalone systems. Level of immersion, perceived whilst using eye-tracking equipment, may be high [27] or low [28], depending on the specific equipment type. Accordingly, proper consideration must be given to the eye-tracking device used in the investigation, to ensure that effective experimental method and data collection is achieved.

The influential issue in the choice of eye-tracking system was the system 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 facilitated 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, these systems were outside our budget. Ultimately we chose the Arrington Research ViewPoint EyeTracker [29] (a Macintosh based system that uses an infrared camera to provide corneal/pupil reflection eye tracking in combination with QuickClamp Hardware. Table II provides a detailed technical specification of ViewPoint Eye-Tracker.

The Arrington Research ViewPoint EyeTracker successfully facilitated streaming video stimulus and allowed appropriate data storage. Eye-tracking data output includes: X coordinate values, Y coordinate values and timing data (a delta time that represents the time {ms} between samples). 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 synchronization between participants, eye-tracking data was sampled at 25 Hz for all participants, corresponding to the maximum experimental frame rate.

IV. EXPERIMENTAL APPROACH

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

A. *Video Content*

The multimedia video clips used were specifically chosen to cover a broad spectrum of infotainment [8]. This range in infotainment content was used to allow for personal preference, and user pre-knowledge, to be cancelled out. 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 (see Fig. 2). The multimedia video clips used varied from those that are informational in nature (such as a news or 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. The duration of video clips used was between 26 and 45 seconds long.

B. *Creating Jitter and Delay Video Material*

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 equaled 2% the number of video frames, which corresponds to one video error every two seconds—the minimum time taken to identify perceptually significant/informative areas in visual stimuli [20], [28], [31]. We appreciate that in a guaranteed memory managed QoS scenario, that time-stamping would not allow 2% jitter or delay, however in a best effort network, using devices with limited memory, such error variation is not unreasonable. Consequently, to simulate accumulated video delay, after every 50 video frames (at 25 frames per second—fps) 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—the

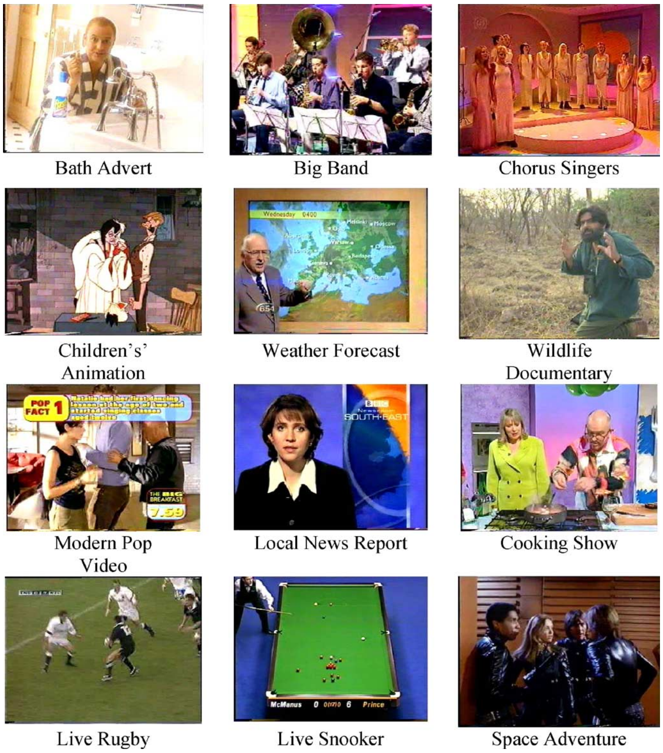

Bath Advert
Big Band
Chorus Singers
Children's' Animation
Weather Forecast
Wildlife Documentary
Modern Pop Video
Local News Report
Cooking Show
Live Rugby
Live Snooker
Space Adventure

Fig. 2. The 12 video clips used in our experiment.

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 (at 25 fps), 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 synchronized with the audio stream. Randomly-sized video skew between 0 and 4 frames were used to ensure variation in jitter, ranging from 0 ms to 160 ms, which represents a maximum skew equal to two times the minimal noticeable synchronization error between video and audio media [8]. Video frame rate variation included 5, 15 and 25 fps video.

C. Experimental Questionnaire

QoP-IA questions characteristically 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, 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. As questions are consistent for all participants any variation is due to other experimental variables.

D. Experimental Variables

Three experimental variables were manipulated in this study, these were: error type, multimedia video frame rate and multimedia video content (Table III).

TABLE III
EXPERIMENTAL VARIABLES AND VALUES

Experimental Variable	Values
Error type	no error (control), delay (2%), jitter (2%)
Frame rate	5, 15, and 25 fps
Video Content	12 clips

TABLE IV
VIDEOS QUALITY TYPES

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)

To consider error variation, original (uncorrupted) as well as delay and jitter video conditions were considered in our experiment. To consider multimedia video frame rates (5, 15 and 25 fps) we introduced video quality types (see Table IV).

Video quality types combine both error type and frame rate, and allows the relationship between each group to be identified. To consider variation multimedia in video content, 12 video clips were considered in our experiment [8]. Since 108 participants were used in our experiment all differences will not be due to the particular preference and pre-knowledge of specific participants.

E. Experimental Methodology

Participant Distribution: 108 users took part in our experiments, with eye tracking being used for all participants. These were divided into three experimental groups, which related to the perceptual impact of control, jitter and delay videos respectively. Participants in each group (36 participants in total) were subdivided into three groups, each containing 12 participants. Sub-groups were used to distinguish the viewing order and frame rate that participants were ultimately going to view multimedia video clips. Participants were aged between 18 and 57 and were taken from a range of different nationalities and backgrounds. All participants spoke English as their first language, or to a degree-level standard, and were computer literate. In each experimental sub-group, e.g. C1, C2, etc., a within-subjects design was used. 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. C3, J3 and D3 sub-group participants all viewed videos with frame-rates as defined by column 'Order 3' (see Table V).

Experimental Setup: To guarantee that experimental conditions remained constant for all control participants, consistent environmental conditions were used. An Arrington Research ViewPoint EyeTracker was used, to extract eye-tracking data, in combination with QuickClamp Hardware. The QuickClamp system is designed to limit head movement and includes chin, nose and forehead rests, whilst supporting the infrared camera.

TABLE V
FRAME-RATE ORDER FOR CONTROL, JITTER AND DELAY SUBGROUPS

Video	Order 1	Order 2	Order 3
Bath Advert	5	15	25
Big Band	25	5	15
Chorus Singers	15	5	25
Children Animation	25	15	5
Weather Forecast	5	25	15
Wildlife Documentary	5	15	25
Modern Pop Video	15	25	5
Local News Report	5	25	15
Cooking Show	15	25	5
Live Rugby	25	5	15
Live Snooker	15	5	25
Space Adventure Series	25	15	5

The position of nose and forehead rests remained constant throughout all experiments (45 cm from the screen). The position of the chin rest and camera were, however, changed depending on the specific facial features of the participant. 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 (70 dB) was used for all participants.

Experimental Process: To ensure that all participants were able to view menu text on the eye-tracker screen without spectacles, each participant was asked to undergo a simple eye-test. 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 throughout the entire visual field. To ensure that the participants did not feel under test conditions, it was made clear that their intelligence was not being tested and that they should not be concerned if they were unable to answer any of the information assimilation questions.

After a brief introduction was given, the ViewPoint eye tracking 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 color 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 throughout the entire visual field. 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 cough.

Once calibration of the eye-tracking system was complete, the appropriate presentation order was loaded and the first video clip was shown. 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 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 verbally, with the answers being noted at the time of asking. Once all participants had successfully completed the experiment eye-tracking data was cleaned, synchronized, and saved in a dedicated data file.

V. RESULTS

A. *Jitter and Delay Impact of on QoP-IA (Information Assimilation)*

An ANOVA (ANalysis Of VAriance) test, with error type (i.e. control, jitter and delay) as the independent variables and QoP-IA as a dependent variable, highlighted that error type has no significant impact on user QoP-IA, which shows that the presence of delay and jitter does not impact a users ability to assimilate information. Moreover, an ANOVA with video quality type (see Table IV) as the independent variable and QoP-IA as a dependent variable showed that quality type does not impact user QoP-IA $\{\mathrm{F}(1,8) = 1.311\ \mathrm{p} = 0.234\}$ (see Fig. 3(a)). In all statistical diagrams confidence intervals (CA) shows the estimated range of values which is likely to -included in the data set $(\mathrm{P} < 0.05)$. Post-hoc Tukey tests are often used to define the significant relationships between factors. Values outside this estimated range are deemed as significant. This result implies that combined network-level quality parameter (jitter and delay) and frame rate variation does not significantly impact user QoP-IA, and shows that a level of error, caused at the network level, does not negatively impact the user's factual understanding of the video content. Although non-significant, this result is important in bandwidth constrained environments as it suggests that users still assimilated factual information independent of a level of network error. Interestingly, an ANOVA with video content as the independent variable and QoP-IA as a dependent variable, showed that video content does significantly impact user QoP-IA $\{\mathrm{F}(1,11) = 12.700\ \mathrm{p} < 0.001\}$ (see Fig. 3(b)), which highlights the importance of infotainment content to the user perceptual experience of multimedia.

B. *Jitter and Delay Impact on QoP-LoQ (Level of Quality)*

An ANOVA with error 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 $\{\mathrm{F}(1,2) = 8.547\ \mathrm{p} < 0.001\}$. Moreover, post-Hoc Tukey-Tests showed a significant difference between the perceived QoP-LoQ for control and jitter $\{\mathrm{p} = 0.001\}$, as well as control and delay videos $\{\mathrm{p} = 0.002\}$. An ANOVA 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 $\{\mathrm{F}(1,8) = 7.706\ \mathrm{p} < 0.001\}$ (See Fig. 4(a)).

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 [6], who showed that the presence of even low amounts of network-level error result 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 multimedia quality should be given priority over and above purely educational applications.

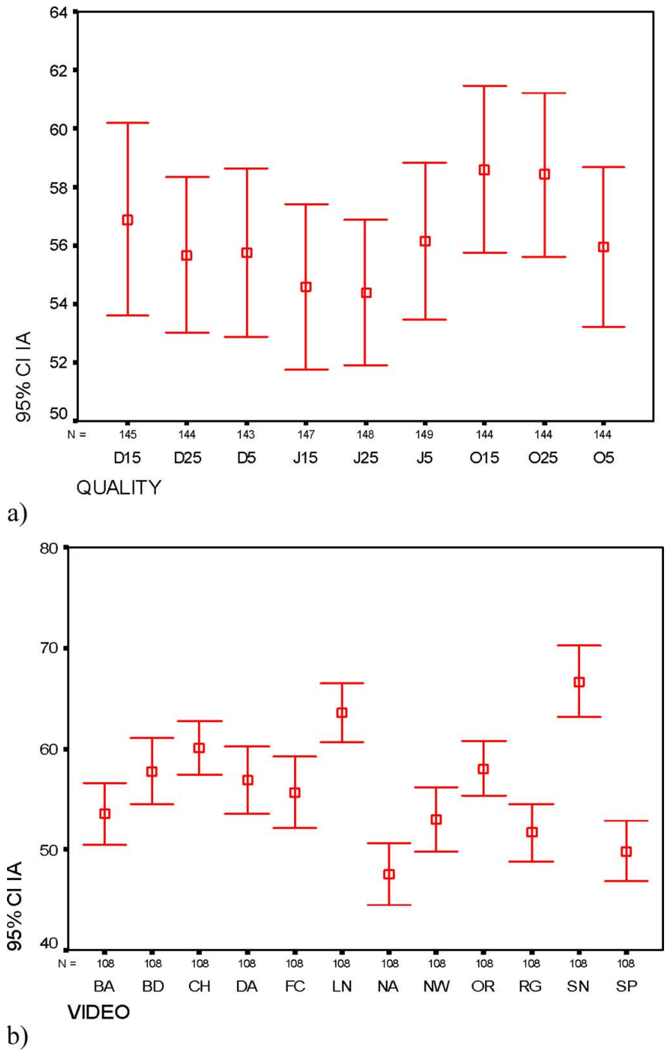


Fig. 3. Impact of (a) quality type (see Table III) and (b) video content on user QoP-IA {Mean and St. Dev.}.

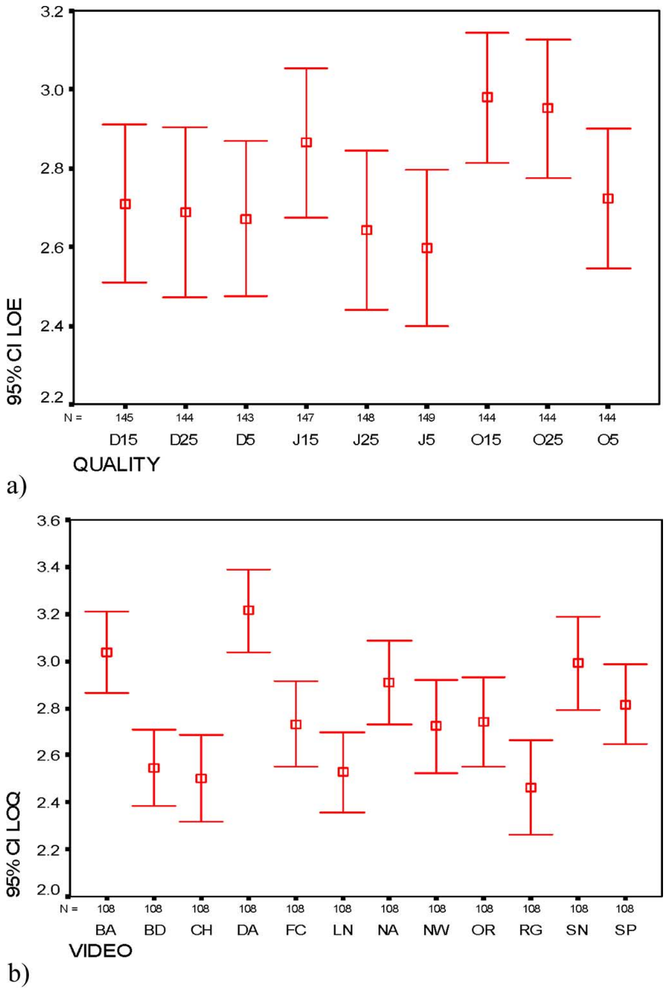


Fig. 4. Impact of (a) quality type and (b) video content user QoP-LoQ {Mean and St. Dev.}.

Interestingly, an ANOVA with video content as the independent variable and QoP-LoQ as a dependent variable, showed that video content significantly affects the user's QoP-LoQ $\{\mathrm{F}(1,11) = 7.085\ \mathrm{p} < 0.001\}$ (See Fig. 4(b)). This is interesting as it suggests that video content (i.e. the content of the information being presented) is of significant importance to the user's perspective of multimedia video quality at the network-level, and consequently should be considered when defining multimedia quality. It also supports the manipulation of information content as a means of improving user QoP-LoQ at the network-level.

C. Jitter and Delay Impact on QoP-LoE (Level of Enjoyment)

A MANOVA (Multiple ANalysis Of VAriance) test, with error type and video quality type as the independent variables and QoP-LoE as a dependent variable, showed QoP-LoE to be significantly impacted both by variation type $\{\mathrm{F}(1,2) = 3.954\ \mathrm{p} = 0.019\}$ and quality type $\{\mathrm{F}(1,8) = 2.221\ \mathrm{p} = 0.024\}$ Fig. 5(a).

Moreover, Post hoc tests show important differences between the control and delay $\{\mathrm{p} = 0.019\}$, and control and jitter $\{\mathrm{p} = 0.037\}$ videos, highlighting that both QoP-S factors (perception of video quality and user enjoyment) are significantly impacted by network-level error.

An ANOVA test with video content as the independent variable and QoP-LoE as a dependent variable, showed that video content significantly impacts user enjoyment $\{\mathrm{F}(1,11) = 8.322, \mathrm{p} < 0.001\}$ (see Fig. 5(b)). It is thus important to note that the type of video being presented is more significant to a user's overall perception of quality (i.e. both the level of information assimilated and user satisfaction) than either variation in presentation frame rate, or the introduction of network-level error (jitter or delay).

D. Jitter and Delay Impact on Eye Gaze

One of the key issues when analysing eye-tracking data is to visualize the areas in which the users gaze is more likely to rest. For temporal dependant data such as video this is often added concern since it introduces a new dimension besides the traditional spatial one. In this section we use two different approaches to analyse eye-tracking data that includes the time domain.

Median Statistical Eye-Gaze Analysis: Eye tracking samples (25 fps) correspond to the maximum frame rate used in

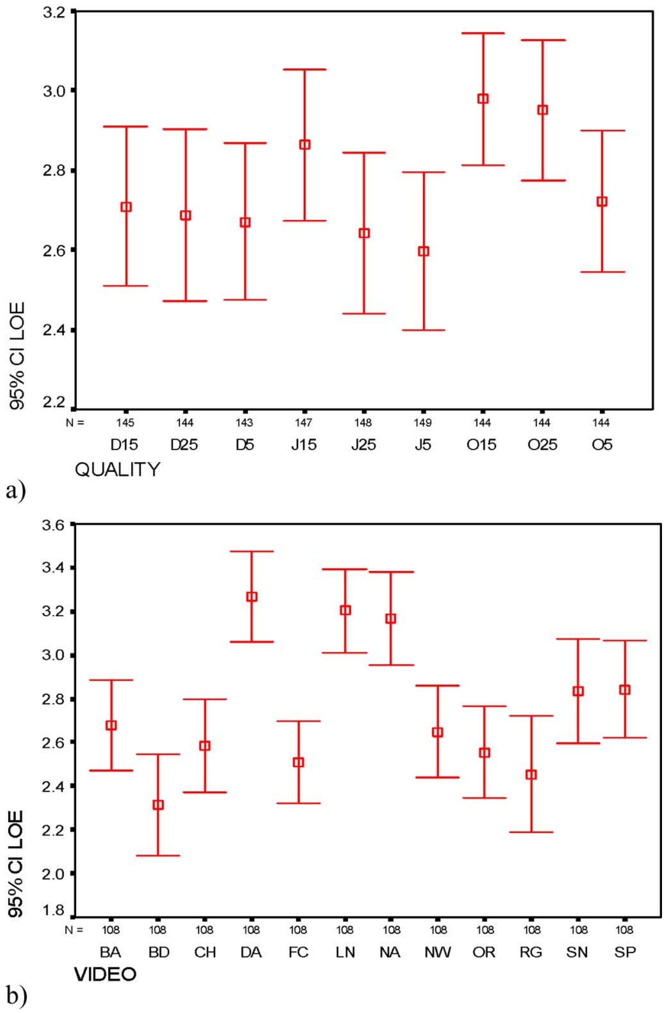


Fig. 5. Impact of (a) quality type and (b) video type on user QoP-LoE {Mean and St. Dev.}.

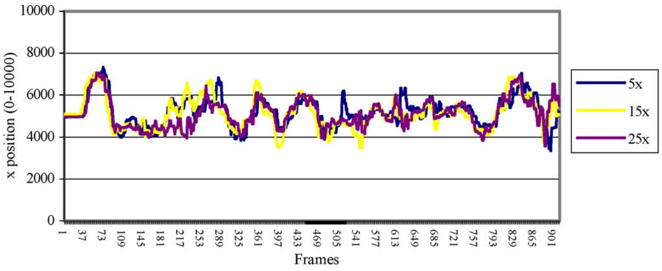


Fig. 6. Space action movie x-coordinate video eye-path.

the experimental material. This facilitates comparison between eye-tracking data and specific video frames. To allow a statistical comparison of eye-gaze data, between frame rates (5, 15 and 25 fps) and quality variation groups, over the duration of each video clip, three (x,y) coordinate points were required for each eye tracker sample, with each sample point relating to a specific frame rate or quality group (see Table IV). As, to the best of our knowledge, no previous eye-tracking data analysis uses statistical comparison across multiple video frames, there was no known precedent for summarizing multiple participant eye-tracking data in this way. Thus, 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 video frame, for each data set (frame rate or quality group)—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, it was considered to be least prone to error values, yet still facilitating statistical analysis. By mapping 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 for each of the defined video quality types (see Table I). The example (see Fig. 6) shows the control x coordinate value for the ‘Space Adventure Series’ clip, which shows a dynamically changing video based around a gun battle [8]. Although, in this specific example, eye fixations tend to return to position 5000 (the center 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 information dimensions: 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 yet facilitate statistical analysis. 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. 5 fps 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 video clip, comparison was also made between control, delay and jitter video groups. These tests were used to establish whether varied frame-rate or network-level error variation (delay and jitter) statistically impacted video eye-paths, i.e. do 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 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-only eye-path frame rate correlation tests show correlation values with significance levels of $p < 0.05$. Therefore, frame rate does not impact user eye-path when video is viewed containing limited delay. Delay videos were involved in 35% of non-correlations, which exist between different error groups (e.g. comparing control and delay), suggesting that the presence of delay in experimental videos causes slight variation in user eye-path from the control. These delay non-correlations were only identified for the big band, chorus singers, and children's animation videos, which implies that the presence of delay is only significant for certain videos.

Jitter: Although the majority of jitter-only eye-path frame rate correlations showed a significant correlation $(p < 0.05)$, there was one noticeable exception—the big band video clip (15 fps/25 fps) $\{r(887) = 0.58, p = 0.85\}$. In addition, jitter videos were involved in 84.6% of non-correlations in video eye-path between different error groups. This implies that the presence of jitter in experimental videos causes considerable variation in user eye-path. Interestingly, jitter non-correlations were also only identified for certain videos: big band, chorus singers, children's animation and the weather forecast videos, which, similar to delay non-correlations, suggests that the use of jitter has more of an effect on specific video.

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 a greater impact than that of delay. Interestingly, disparity only happens to four out of twelve videos (Big band, Chorus singers, Children's animation and the Weather forecast). Although this is most probably due to the existence of multiple Regions of Interest, this conclusion cannot be made at this time.

Fixation Maps: Fixation maps were first introduced by Wooding [32], 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 digitized 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 facilitate the visual representation of multi-user eye-tracking data. Essentially, a fixation map spatially plots the areas if an image/video frame that possess the greatest number of user fixations (the coordinated focusing of eyes on a particular region). Fixation maps are better when considered as a terrain or a landscape and are color coded. So, in the case of a greyscale map, a pixel value of 0 would represent no user attentive interest, while 255 would mean maximum attentive user interest at that particular location. In our work fixation maps, were found to be particularly useful because they highlighted a video frame's Regions of Interest (RoIs), i.e. the areas that users are most likely to rest their gaze on.

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 not only allowed RoIs to be mapped in the form of an image, but also allow the difference between two data sets to be determined. 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 in user focus as a result of error type (see Fig. 7(e) and Fig. 7(f)). By analysing the average pixel value for consecutive difference fixation maps, we can identify specific sections of videos where a higher level of user video eye-path variation exists, i.e. a relative greyscale variation as a result of delay and jitter.

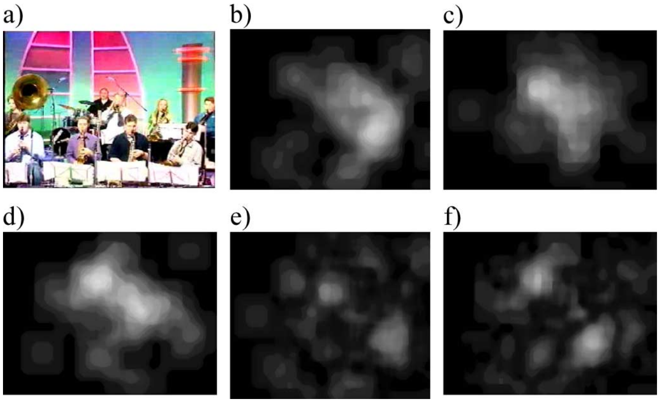


Fig. 7. (a) Original frame; (b) control fixation map; (c) delay fixation map; (d) jitter fixation map; (e) pixel difference between control and delay RoI areas; (f) pixel difference between control and jitter RoI areas.

Analysis suggests that high levels of disparity in user eye-path occur for two reasons: i) when no single/obvious point of focus exists, causing a conflict of user attention; or ii) when the point of attention changes dramatically, 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.

It is important to note that dynamic video content does not negatively impact user regions 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 also occurs, e.g. the man wiping the bath with the sponge.

Previously non-correlations of video eye-path were found between different error groups. This finding suggests that the presence of delay and jitter introduces variation in user eye-path. Interestingly, these non-correlations only existed for videos, which, for the majority of the video, do not have a single/obvious point of focus. Instead multiple conflicting points of focus exist, which ultimately cause variation in user eye-path, as a result of delay and jitter. Moreover, the level of variance in user focus changes as a result of fast scene changes or multi person dialogue, i.e. does not ensure smooth pursuit eye movement, thus resulting in problems with identifying and tracking important regions-of-interest.

VI. Conclusions

In this paper we presented the results of a study that examined the user's perception of multimedia quality, when impacted by delay and jitter. Additionally, eye-tracking was incorporated as a tool enabling continuous monitoring of attention; thus providing a better understanding of the role that the user plays in the reception, analysis and finally the synthesis of multimedia data.

Results showed that whilst, the type of multimedia video content impacts user QoP-IA, there was no statistically significant difference in the level of QoP-IA, due to error type (i.e. control, jitter and delay). The latter demonstrates that delay and jitter does not negatively impact information assimilation; this is an important result for broadcast and distributed educational applications, as it shows that QoS degradation does not impact the users' ultimate understanding of the video content.

Error type was, however, found to significantly affect both user QoP-LoQ and QoP-LoE, thus implying that not only can a user distinguish between a video presentation with and without error, but the presence of error impacts the user's overall level of enjoyment. This finding, whilst at first reading might seem common sense, does go against the grain of previous research [6], [7], [13], [14], which has highlighted user perceptual tolerance to QoS artifacts when content is viewed purely for entertainment purposes. However, our results are of particular importance to broadcast and bandwidth-constrained environment, for they show that when such content is viewed for infotainment purposes, the perceptual tolerance of users to such artifacts is heightened. Thus, in such cases, it is also critical that network QoS variation be minimized in order to minimize impact on user QoP-S. Moreover, findings support [3], who observed that degradation of network level QoS has a greater influence on a subjects' uptake of emotive content than on their uptake of factual content. Lastly, results showed that variation in video eye path occurs when: i) no single or obvious point of focus exists; or ii) when the point of attention changes dramatically.

Our work has provided a better understanding of what is, in a multimedia context, user perceived quality and how this relates to user eye gaze patterns. It does, however, raise numerous issues, which require further research. Firstly, if video content is more significant to a user's definition of multimedia quality than the occurrence of network-level error, could adaptation of video content be used to increase user quality perception? Secondly, could adaptive communication systems be constructed which take into account user perceptual tolerances to network level degradations—both issues represent areas of future endeavor upon which we shall be concentrating our attention.

REFERENCES

[1] M. Claypool and J. Tanner, "The effects of jitter on the perceptual quality of video," in *ACM Multimedia '99 (Part 2)*, Orlando, FL, 1999, pp. 115–118.

[2] G.-M. Muntean, P. Perry, and L. Murphy, "Subjective assessment of the quality-oriented adaptive scheme," *IEEE Trans. Broadcasting*, vol. 51, no. 3, pp. 276–286, 2005.

[3] R. Procter, M. Hartswood, A. McKinlay, and S. Gallacher, "An investigation of the influence of network quality of service on the effectiveness of multimedia communication," in *Proc. of the International ACM SIGGROUP Conference on Supporting Group Work*, New York, USA, 1999, pp. 160–168, ACM.

[4] Y. Wang, M. Claypool, and Z. Zuo, "An empirical study of RealVideo performance across the internet," in *Proc. of the First ACM SIGCOMM Workshop on Internet Assessment*, New York, USA, 2001, pp. 295–309, ACM Press.

[5] G. Ghinea, "Quality of Perception—An Essential Facet of Multimedia Communications," , Department of Computer Science, The University of Reading, , UK, 2000, Submitted for the Degree of Doctor of Philosophy.

[6] D. Wijesekera and J. Srivastava, "Quality of Service (QoS) metrics for continuous media," *Multimedia Tools Applications*, vol. 3, no. 1, pp. 127–166, 1996.

[7] J. A. Kawalek, "User perspective for QoS management," in *Proc. of the QoS Workshop Aligned With the 3rd Internationals Conference on Intelligence in Broadband Services and Network (IS&N 95)*, Crete, Greece, 1995.

[8] R. Steinmetz, "Human perception of jitter and media synchronization," *IEEE Journal on Selected Areas in Communications*, vol. 14, no. 1, pp. 61–72, 1996.

[9] O. Verscheure, P. Frossard, and M. Hamdi, "User-oriented analysis in MPEG-2 video delivery," *Journal of Real-Time Imaging*, vol. 5, no. 5, pp. 305–314, October 1999, 1999.

[10] W. Dapeng, Y. T Hou, W. Zhu, Y. Q. Zhang, and J. M. Peha, "Streaming video over the internet: Approaches and directions," *IEEE Trans. Circuits Systems for Video Tech.*, vol. 11, no. 3, pp. 282–300, 2001.

[11] G. M. Muntean, P. Perry, and L. Murphy, "A new adaptive multimedia streaming systems for all-IP multi-service networks," *IEEE Trans. Broadcasting*, vol. 50, no. 1, pp. 1–10, March 2004.

[12] 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.

[13] 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, Scotland, 1997, pp. 189–194.

[14] A. Bouch, G. Wilson, and M. A. Sasse, "A 3-dimensional approach to assessing end-user quality of service," in *Proc. of the London Communications Symposium*, 2001, pp. 47–50.

[15] P. Faraday and A. Sutcliffe, "Authoring animated web pages using "contact points"," in *Proc. of the SIGCHI Conference on Human Factors in Computing Systems: The CHI Is the Limit*, Pennsylvania, US, May 1999, pp. 458–465.

[16] P. Faraday and A. Sutcliffe, "Making contact points between text and images," in *Proc. of ACM Multimedia '98*, Bristol, UK, 1998, pp. 29–37.

[17] L. Kaufman and W. Richards, "Spontaneous fixation tendencies for visual forms," *Perception and Psychophysics*, vol. 5, pp. 85–88, 1969.

[18] SMPTE Psychophysics Subcommittee White Paper, G. Elias, G. Sherwin, and J. Wise, "Eye movements while viewing NTSC format television," 1984.

[19] B. L. Cole and P. K. Hughes, "Drivers don't search: They notice," in *Visual Search*, D. Brogan, Ed. Amsterdam: Elsevier Science Publishers, 1990, pp. 407–417.

[20] A. L. Yarbus, "Eye movement and vision," in *Trans. B. Haigh.*. New York: Plenum Press, 1967.

[21] A. Gale, "Human response to visual stimuli," in *The Perception of Visual Information*, W. Hendee and P. Wells, Eds. : Springer-Verlag, 1997, pp. 127–147.

[22] J. Senders, "Distribution of attention in static and dynamic scenes," in *Proc. SPIE*, San Jose, 1992, vol. 3016, pp. 186–194.

[23] P. Isokoski, "Text input methods for eye trackers using off-screen targets," in *Proc. of the Symposium on Eye Tracking Research and Applications 2000*, Palm Beach Gardens, Florida, United States, 2000, pp. 15–21.

[24] E. Reingold and L. C. Loschky, "Reduced saliency of peripheral targets in gaze-contingent multi-resolutional displays: Blended versus sharp boundary windows," in *Proc. of the Symposium on ETRA 2002: Eye Tracking Research and Applications Symposium 2002*, New Orleans, Louisiana, 2002, pp. 89–93.

[25] J. H. Goldberg, M. J. Stimson, M. Lewenstein, N. Scott, and A. M. Wichansky, "Eye tracking in web search tasks: Design implications," in *Proc. of the Symposium on ETRA 2002: Eye Tracking Research & Applications Symposium 2002*, New Orleans, Louisiana, 2002, pp. 51–58.

[26] M. Sodhi, B. Reimer, J. L. Cohen, E. Vastenburg, R. Kaars, and S. Kirschenbaum, "Onroad driver eye movement tracking using head-mounted devices," in *Proc. of the Symposium on ETRA 2002: Eye Tracking Research & Applications Symposium 2002*, New Orleans, Louisiana, 2002, pp. 61–68.

[27] M. M. Hayhoe, D. H. Ballard, J. Triesch, P. Aivar, and B. Sullivan, "Vision in natural and virtual environments," in *Proceedings of the Symposium on ETRA 2002: Eye Tracking Research & Applications Symposium 2002*, New Orleans, Louisiana, 2002, pp. 7–13.

[28] T. Partala, M. Jokiniemi, and V. Surakk, "Responses to emotionally provocative stimuli," in *Proc. of the Symposium on Eye Tracking Research and Applications 2000*, Palm Beach Gardens, Florida, United States, 2000, pp. 123–128.

[29] Arrington Research, "Arrington Research Eyetracker," Dec. 18, 2006 [Online]. Available: http://www.arringtonresearch.com

[30] 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.

[31] J. F. Mackworth and A. J. Morandi, "The gaze selects informative details within pictures," *Perception and Psychophysics*, vol. 2, pp. 547–552, 1967.

[32] D. S. Wooding, "Fixation maps: Quantifying eye-movement traces," in *Proc. of the Symposium on ETRA 2002: Eye Tracking Research & Applications Symposium 2002*, New Orleans, Louisiana, 2002, pp. 31–36.

Stephen R. Gulliver (M'02) received the B.Eng. (Hons) degree in microelectronics (1999), an M.Sc. degree in distributed information systems (2001), and a PhD from Brunel University (2004), United Kingdom. He is a Lecturer in the Department of Information Systems and Computing at Brunel University. His research interests Human Factors, including: perceptual aspects of multimedia, 3D and Virtual Reality 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 Computer Science from the University of Reading, United Kingdom, in 2000. He is a Senior 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.