A Perceptual Comparison of Empirical and Predictive Region-of-Interest Video

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

***Abstract*—When viewing multimedia presentations, a user only attends to a relatively small part of the video display at any one point in time. By shifting allocation of bandwidth from peripheral areas to those locations where a user's gaze is more likely to rest, attentive displays can be produced. Attentive displays aim to reduce resource requirements while minimizing negative user perception—understood in this paper as not only a user's ability to assimilate and understand information but also his/her subjective satisfaction with the video content. This paper introduces and discusses a perceptual comparison between two region-of-interest display (RoID) adaptation techniques. A RoID is an attentive display where bandwidth has been preallocated around measured or highly probable areas of user gaze. In this paper, video content was manipulated using two sources of data: empirical measured data (captured using eye-tracking technology) and predictive data (calculated from the physical characteristics of the video data). Results show that display adaptation causes significant variation in users' understanding of specific multimedia content. Interestingly, RoID adaptation and the type of video being presented both affect user perception of video quality. Moreover, the use of frame rates less than 15 frames per second, for any video adaptation technique, caused a significant reduction in user perceived quality, suggesting that although users are aware of video quality reduction, it does impact level of information assimilation and understanding. Results also highlight that user level of enjoyment is significantly affected by the type of video yet is not as affected by the quality or type of video adaptation—an interesting implication in the field of entertainment.**

***Index Terms*—Attentive displays, eye tracking, perceptual quality, region of interest (RoI).**

I. Introduction

VISUAL information is computationally intense, yet due to improved rendering hardware, high-quality graphical displays are now commonplace. With increasing screen sizes and screen resolutions, the user has a growing expectation of what defines high-quality video, particularly when transmitted over bandwidth-constrained environments. Nonetheless, most of the resources used to produce large high-resolution displays are wasted, as the user never actually looks at the whole screen at one point in time: Ocular physiology limits the range of high acuity to approximately 2° of the visual field, which is equivalent to approximately the width of a thumbnail at arm's length or 2 cm at a typical reading distance of 30 cm [1]. It is interesting to note that if the human visual system can only process detailed information within an area at the center of vision (with rapid acuity drop-off in peripheral areas [21]), the point of user attention, if effectively monitored or predicted, can be used to manage nonuniform allocation of bandwidth. Such displays are known as attentive displays. Current attentive displays form the focus of a wide range of applications (see, for instance, [5], [20], [22], and [29]), including reading, perception of image and video scenes, virtual reality, computer game animation, art creation and analysis, and visual search studies [1].

Attentive displays have traditionally been synonymous with the field of gaze-contingent displays (GCDs). GCDs select the region of interest (RoI)—the area on the screen where the user is currently looking—by actively tracking the viewer's eyes in real time and maintaining a high level of detail at the point of gaze. Early attempts at attentive displays suffered from limited display sizes, noticeable quality edges, and limited control of resolution [32]. Increased screen and resolution sizes, as well as the falling cost of eye-tracking equipment, have all led to increased interest in GCDs; however, such displays still suffer from considerable practical and application issues. GCDs necessitate an eye-tracker device with high sample rates, with corresponding computational, technical, and cost implications, to ensure appropriate refresh rates (4–15 ms, depending on window sizes [19], [26]). In addition, GCDs can only be used in single-user environments, since the eye-tracking device is only capable of tracking the gaze of a single user. To overcome these limitations, we propose the concept of *region-of-interest displays* (RoIDs). RoIDs use predefined RoI areas to adapt the video quality such that resource allocation is biased in attentive areas. RoI areas can be obtained from either prior empirically collected eye-tracking data or from predicted analysis of video content.

In this paper, we present the results of a study that compares the perceptual impact of adapting RoIDs with either empirically obtained eye-tracking data or computationally defined predictive RoI data.

II. Attentive Processes and Displays

A. Attentive Processes

Light reflected from objects in the visual field enters the eye and passes through the lens, which projects an inverted image of an object onto the retina at the back of the eye. The retina itself consists of approximately 127 million light-sensitive cells, roughly split as follows: 120 million rods (sensitive to brightness) and 7 million cones (sensitive to color).

Manuscript received July 21, 2006; revised May 22, 2007, November 9, 2007, and April 8, 2008. Current version published June 19, 2009. This paper was recommended by Associate Editor L. Rothrock.

S. R. Gulliver was with the Department of Information Systems and Computing, School of Information Systems, Computing and Mathematics, Brunel University, Uxbridge UB8 3PH, U.K. He is now with the Informatics Research Centre, University of Reading, Reading RG6 6WB, U.K.

G. Ghinea is with the Department of Information Systems and Computing, School of Information Systems, Computing and Mathematics, Brunel University, Uxbridge UB8 3PH, U.K.

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/TSMCA.2009.2019893

Both cones and rods are unevenly distributed across the retina. In particular, 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 center of the retina, in a circular area called *macula lutea*. Within this area, there is a depression called the *fovea*, which consists almost entirely of cones, and it is through this area of high acuity, extending over just 2° of the visual field, that humans make their detailed observations of the world. High acuity color vision therefore relies on cone receptors located in the fovea. Movements of the eye, head, and body are used to bring RoIs into the visual path at the center of the fovea. This movement between items within the stationary field, the eye field, and the head field is determined by visual attention [30].

The eye naturally selects/fixates on areas that are likely to be most informative [16]. In between fixations, eye behavior is characterized by saccades (rapid eye movements between regions of informative interest). The process of visual attention is thus broken into two sequential stages: the preattentive stage and the selection stage [34]. In the preattentive stage, information is processed from the whole visual field in parallel and subconsciously defines objects from visual primitives, including motion [12], contrast [7], [37], size [7], [23], shape [8], color [2], [23], brightness [17], line edges [13], [37], and orientation [18]. It is the preattentive stage that determines RoIs within the visual field (defining important visual cues), and based on this preattentive mapping, the selection stage performs high-level serial processes, dependent on high-level search criteria, including location [5], foreground/background [2], [37], and introduction of people and context [8], [37]. When items pass from the preattentive stage to the selection stage, these items are considered as selected. Four distinct looking states have been defined, which summarize the cognitive state of the user [15]: *spontaneous looking* (when a subject is not actively looking for, or thinking about, any specific object), *task-relevant looking* (when a subject is performing a specific task, such as reading), *orientation of thought looking* (where eye movements represent a general orientation toward the object of thought), and *intentional manipulatory looking* (where subjects consciously control their direction of looking).

We are particularly interested in eye movements, since they offer insights into visual perception, as well as the associated attention mechanisms and cognitive processes. Moreover, interpretation of eye-movement data is 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 [14]. Last but not least, as mentioned earlier, the eye naturally focuses on areas that are more likely to be informative [21].

B. Adaptive Attentive Displays

While measuring eye movements, Stelmach *et al.* [33] showed that eye movements during television viewing are not idiosyncratic to a specific viewer but that the direction of gaze is highly correlated among viewers. This view is supported by the study in [11] and supports the use of attentive displays, particularly in bandwidth-constrained environments, with the aim of minimizing the use of bandwidth resource, while limiting negative impact on user perception. As mentioned in Section I, two main approaches have been developed to implement attentive displays: GCDs (the traditional approach) and RoID systems (which are being introduced in this paper). GCD systems facilitate real-time user-specific attention-based rendering. However, they require an eye-tracker device with high sample rates and only support a single viewer. Although not "real-time," RoIDs are defined in the following: 1) RoIDs can be achieved at a fraction of the cost of GCDs; 2) can be developed to accommodate multiple users; and 3) facilitate distributed bandwidth savings since video can be preencoded, thus reducing bandwidth needs prior to transmission. Moreover, RoIDs do not require each user to possess specialized eye-tracking or video-processing hardware and can therefore easily be integrated with current display systems, ultimately making RoIDs more commercially attractive, particularly in bandwidth-constrained environments.

Studies looking at the benefits and disadvantages of using GCDs have shown a range of results. Duchowski and Coltekin [5], also providing a comprehensive introduction to the area of GCDs (covering topics such as medical applications of GCDs and perceptually lossless GCDs), present an elegant pixel shader algorithm showing that real-time GCD processing can be used on both still image and video content. This implication is that GCDs can be implemented for real-time video feeds and that they support the use of video in bandwidth-constrained environments. In related work, Loschky and Wolverton [20] considered the issue of continued perceptual disruptions in GCDs—specifically examining perceptually acceptable update delays in multiresolution displays. This research suggests that GCDs, although useable, could introduce considerable perceptual distraction that can interrupt normal attentive processes. In parallel research, Reingold and Loschky found that when they adapted a high-resolution window at the point of gaze and degraded resolution in peripheral areas, participants had longer initial saccadic latencies in peripheral areas (the time taken to identify a visual target) than when a low resolution was uniformly displayed across the whole display window [27]. This implies that the use of degraded peripheral resolution can lead to longer search times and therefore impact user preattentive processes. Moreover, Loschky and McConkie found, in support of earlier studies [31], that if degradation is increased in peripheral areas, the size of the adapted high-resolution window at the point of gaze also needs to be increased, if the user's level of performance is to be maintained [19]. However, this increase in the high-resolution window cancels out any bonus of peripheral degradation and limits any gain of using a GCD system.

The use of high-resolution/high-quality regions is central to most attention-based displays, and therefore, Loschky and McConkie place into doubt the effectiveness of windowed gaze-contingent multiresolution displays [19]. Reingold and Loschky [27] also suggested that reduced reactions may be due to participants being distracted by boundary edges (a change in visual resolution). Consequently, Reingold *et al.* compared sharp and blended resolution boundary conditions to identify whether increased saccadic delays were due to boundary lines [28]. Three conditions were used: 1) the no-window condition (where all of the images were blurred or of lower quality); 2) a 12° window with no blending; and 3) a 12° window with a 3° wide region of blending. However, results showed that the

first condition did, indeed, produce shorter mean initial saccades, supporting the work of Czerwinski *et al.* [4] who stated that a wider field of view can lead to increase performance in productivity. Interestingly, Reingold *et al.* found no difference in a user's ability to identify visual errors, as a result of a "window" edge being either sharp or softened. This is in distinct conflict with previous work [35], which indicated that the blending regions were vital to the perceptual quality of the display.

Arguments presented in [19] and [26] place considerable doubt on whether GCD displays provide any perceptual advantage, particularly in low-bandwidth environments. Interestingly, Osberger *et al.* demonstrate a technique for controlling adaptive quantization processes in an MPEG encoder, based on frame-based importance maps (IMs) [25]—a form of RoID. IMs are produced using segmented images, which are analyzed using five factors, namely, contrast, size, shape, location, and background importance. Lower quantization was assigned to visually important regions, while areas that were classified as being of low visual importance were more harshly quantized. This method was evaluated on a wide variety of images with results indicating that IMs significantly correlate well with human perception of visually important regions [24] and can be used to support a reduction of bandwidth without a reduction in user perception of quality.

In summary, different attentive display implementation approaches (GCDs or RoIDs), as well as the use of different adaptation data (empirical and predictive RoI data), appear to have a significant impact on users' reaction and ability to notice video presentation errors. Interestingly, no studies to our knowledge have considered this issue. Accordingly, this paper measures the perceptual impact of using attentive RoIDs by comparing the perceptual impact of empirical eye-gaze and predictive content-dependent data.

III. QoP

In order to explore the human side of the multimedia experience, we have used the quality-of-perception (QoP) metric. QoP captures the multimedia infotainment duality and encompasses not only a user's satisfaction with the quality of multimedia presentations (denoted by QoP-S) but also his or her ability to understand and assimilate the informational content of multimedia (denoted by QoP-IA). QoP-S is subjective in nature and, in this paper, consists of two component parts: QoP-LoE (the user's Level of Enjoyment while viewing the multimedia content) and QoP-LoQ (the user's judgement concerning the Level of Quality of service provision).

A. QoP Versus QoS

In a distributed setting, the quality of digital multimedia has traditionally been measured using just quality-of-service (QoS) technical parameters. Although measurable, such objective parameters disregard the user's perception of what defines multimedia quality. Due to the multidimensional nature of multimedia, however, it is impossible to rely purely 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.

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 measured.

1) Measuring IA and Understanding (QoP-IA): In our approach, QoP-IA was expressed as a percentage measure, which reflected a user's level of understanding and information assimilation (IA), from visualized 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. 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, e.g., pertaining to the activity of lions in the documentary clip.
- A Information which is presented in the audio stream, e.g., the audio content of the news.
- T Textual information contained in the video window, e.g., the newscaster's name in a caption window.

For each clip, the number of questions targeting a particular information source was roughly proportional to the importance of that source (as given by the weighting of Table I) in the context of the clip. All IA questions must have unambiguous answers, making it possible to determine if a participant had answered them correctly or not. Since, in our experiments, questions can only be answered if information is understood and assimilated from specific information sources, it is possible to determine the percentage of correctly answered questions that relate to the different information sources within a specific multimedia video clip. Thus, by calculating the percentage of correctly answered questions from different information sources, it was possible to generalize from which information sources participants absorbed the most information. By using these data, it is possible to determine and compare, over a range of different multimedia content, potential differences that might exist in a user's understanding of the informational content of the multimedia video, namely, QoP-IA.

2) Measuring Subjective LoQ (QoP-LoQ): The first component part of QoP-S is the users' subjective Level of video Quality of service. In order to measure this, users were asked to indicate, on a scale of 0–5, how much they judged, independent of the subject matter, the presentation quality of a multimedia that they had just seen (with scores of 0 and 5, respectively, representing "no" and "absolute" user satisfaction with the multimedia presentation quality).

3) Measuring Subjective LoE (QoP-LoE): The other component part of QoP-S is the subjective LoE (QoP-LoE) experienced by a user when watching a multimedia presentation. To measure QoP-LoE, the user was asked to express, on a scale of 0–5, how much they enjoyed the video presentation (with scores of 0 and 5, respectively, representing "no" and "absolute" user satisfaction with the multimedia video presentation).

IV. Experimental Methodology

The aim of the study presented in this paper was to measure and compare the impact that empirical and predicted RoI data

TABLE I
CLIP DESCRIPTIONS—CHARACTERISTIC WEIGHTINGS, PREVIOUSLY DEFINED BY GHINEA AND THOMAS [9], DESCRIBE THE VIDEO DYNAMIC (D), VIDEO (V), AUDIO (A), AND TEXTUAL (T) COMPONENTS; SCORES OF 0–2, RESPECTIVELY, REPRESENT LOW, MEDIUM, AND HIGH

		D	V	A	T
Band (BD)	A high school band playing a jazz tune against a multicolored and changing background of lights. No textual information is included.	1	1	2	0
Bath Commercial (BA)	An advertisement for a bathroom cleaner, with information transmitted via the narrator, both audio and visually by the couple being shown in the commercial, and textually, through a slogan.	1	2	2	1
Chorus (CH)	A chorus comprising eleven members performing mediaeval Latin music. No textual information is included.	0	1	2	0
Oregano Cooking (OR)	A familiar television cooking show. Although mainly static, considerable information is being passed to the viewer through the audio dialogue and visually, through the presentation of ingredients. No textual information is included.	0	2	2	0
Animation (DA)	An animated disagreement between two main characters. The clip includes several subtle nuances, e.g. the correspondence between the stormy weather and the argument. No textual information is included.	1	2	1	0
Weather Forecast (FC)	A clip concerning forthcoming weather in Europe and the U.K. Information is presented visually (through the weather maps), textually (information including temperatures, visibility in foggy areas) and through oral presentation of the forecaster.	0	2	2	2
Indian Lions (LN)	A documentary about lions in India. Both audio and video streams are important. No textual information is included.	1	2	2	0
Natalie's Pop Music (NA)	This clip is characterized by the unusual importance of the textual component, which details facts about the singer's life. From a visual viewpoint it is characterized by the fact that the clip was shot from a single camera position.	1	2	2	2
News (NW)	Two stories are presented: one purely verbal, whilst the other has supporting video footage. Rudimentary textual information (channel name, newscaster's name) is displayed at various stages.	0	2	2	1
Rugby (RG)	A test match between England and New Zealand, including the scoring of a try. Essential textual information (i.e. the teams and the score) is displayed in the upper left corner of the screen. The clip is characterized by great dynamism.	2	2	1	0
Snooker (SN)	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.	0	1	1	2
Space (SP)	An action scene from a popular science fiction series including rapid scene changes, accompanying visual effects (explosions).	2	2	1	0

have on user perception of RoIDs. Accordingly, the study was separated into three distinct phases.

1) The first phase comprised an experiment which was run with a control group visualizing multimedia video content without any gaze-contingent adaptation. Participants' eye-gaze location was recorded, and the data used to determine relevant RoIs for each frame of the multimedia clips were viewed.
2) The second phase involved the implementation of *empirical* and *predictive* RoIDs for the multimedia content used in our study. The former RoID used data obtained from the first phase of the study, while the latter was based on automated analysis of video content.
3) The third and last phase of our study comprised measuring and comparing the perceptual impact of empirical and predictive RoIDs using the QoP metric.

A. Multimedia Content

To ensure consistency, identical multimedia video clips were used throughout the experimental process. The multimedia video clips were chosen to cover a broad spectrum of infotainment subject matter. Multimedia video clips vary in nature from those that are informational (such as a news/weather forecast) to ones that are usually viewed purely for entertainment purposes (such as an animation, a music clip, or a sports event, as detailed in Table I).

B. Phase 1: Control Experiment

1) Participants: To obtain eye-tracking data, yet ensure that participants had a consistent type of looking for both phases 1 and 3 of our study, our control experiment incorporated the QoP experimental process. Thirty-six participants were evenly divided into three experimental groups, used to distinguish the viewing order and frame rate at which participants viewed multimedia video clips. Participants were aged between 21 and 55 and volunteered to take part in the study. They were recruited from the authors' circle of academic, professional, and personal contacts, specifically to represent a range of different nationalities and backgrounds. All participants spoke English as either their first language or to a degree-level standard. All participants were computer literate. In our study, this category defined those users who professed to being proficient at using Internet/Web applications as well as standard desktop applications, such as word processors and spreadsheets.

2) Setup: To ensure consistent participant looking, it was vital that the process used to obtain eye-tracking data, to enable an empirical RoID, was the same as the process used during perceptual experimentation. Inconsistent use of method could lead to varied participant looking, which, in turn, might result in adaptation of non-RoI areas. Consequently, the method described in Section III-B was also used when we collected empirical data.

In our control experiment, a within-subjects design was chosen. Thus, each participant viewed four video clips at 5 frames

per second (fps), four at 15 fps, and four at 25 fps, in order to view content with a wide spectrum of QoS. Moreover, in order to counteract any possible order effects, the video clips were shown in a number of order and frame-rate combinations.

To guarantee that experimental conditions remained constant for all control participants, consistent environmental conditions were used. An Arrington Research, Power Mac G3 (9.2) infrared camera-based pupil tracking, ViewPoint EyeTracker was used to extract eye-tracking data, in combination with QuickClamp Hardware. The QuickClamp system is designed to limit head movement, including chin, nose, and forehead rests. Consequently, the position of nose and forehead rests remained constant throughout all experiments (45 cm from the screen). The positions 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.

3) Experimental Process: To ensure that all participants were able to view menu text on the eye-tracker screen without spectacles, each was asked to take part in 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. Users were given an introduction to the experiment. They were then 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 shades of participants' facial features varied considerably, time was taken to adjust the chin rest, infrared capture camera, and software settings to ensure that pupil fix was maintained throughout the entire visual field. Once the configuration setup was complete, automatic calibration was made using a full-screen stimulus window.

When calibration was complete, the appropriate presentation order was loaded, and the presentation state was incremented, which started the first video clip. 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 shown. 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. The participant was asked questions orally, and the answers were all noted at the time of asking.

C. Phase 2: Implementation of Empirical and Predictive RoIDs

1) Extracting Eye-Tracking Data: Empirical RoIDs were determined using data obtained in phase 1 of our study. Eye-tracking data samples contained X values, Y values, and timing data (synchronized during data cleaning for a specific frame). X- and Y-coordinate values (in the range of 0–10 000) were defined automatically by the ViewPoint EyeTracker system

Fig. 1. Eye-based RoI areas.

and, respectively, represented the minimum and 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 (10 000, 10 000). In order to simplify data comparison between participant sets, eye-tracking data were sampled at 25 Hz for all clips used as part of our experiments, corresponding to the maximum frame rate being displayed.

For each video frame, empirical eye-based RoI data were extracted by taking the fixation data from each participant $[XY]$ and identifying a RoI square ($\pm 4^\circ$ of the visual field) that was centered around the coordinate pair location (Fig. 1). $\pm 4^\circ$ of the visual field was used to ensure that the area of participant focus was always covered at a higher frame rate [19] while accounting for the potential tolerance in the used eye tracker. Accordingly, 36 RoI squares were recorded for each of the video frames contained in the multimedia clips, facilitating the implementation of an empirical RoID. RoID squares were exported to a RoID script, which contains all RoI squares for each frame of each multimedia video clip.

2) Extracting Predictive RoI Data: In our study, predictive RoI data were obtained through automated analysis of video content. This was a two-stage process, the first of which computed *primitive images*, with the second extracting RoIs for each primitive image. Accordingly, the primitive images considered in our study were based on the visual primitives of color contrast, edges, and movement, identified in the literature [2], [12], [13], [23], [37] as being perceptually relevant.

1) Color contrast images: A 640 $\times$ 480 pixel image was extracted from each video frame [see Fig. 2(a)], for all 12 video clips described in Section IV-A. These color images were subsequently used to calculate the areas of frames that had a high level of color contrast.
2) Edge images: Edges characterize boundaries and are therefore fundamentally important in image processing. Most edge-detection methods work on the assumption that this is a very steep gradient in the image; accordingly, by using a weighted mask, it was possible to detect edges across a number of pixel values [see Fig. 2(b)].
3) Movement images: The pixel difference between frame N and $N+1$ determined the level and location of movement in subsequent video frames. Software was developed to clearly identify the pixel difference between two subsequent frames [see Fig. 2(c)].

A distribution of the RGB pixel values was made for each color, edge, and movement image. This allowed mean pixel and standard deviation values to be determined for each image. By combining image data, a mean pixel and standard deviation value was calculated for each of the 12 video clips. Important regions of color, edges, and movement can be identified by assuming the following: For color, an abnormal distribution of

Fig. 2. Primitive Images. (a) Original video frame. (b) Edge detection in video frame. (c) Motion detection in video frame.

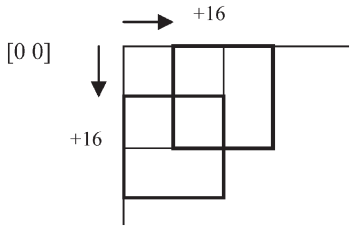

Fig. 3. Overlapping pixel squares.

Fig. 4. Content-dependent RoI areas.

color suggests an area of contrast or an abnormal color; for edges, an abnormal average pixel value suggests a greater level of black lines, i.e., edges; and for movement, a higher than average pixel value suggests a greater variation level of pixel values between frames, i.e., movement.

To calculate significant areas of specific frames, the color, edge, and movement images for each image were split into overlapping 32 $\times$ 32 pixel squares (Fig. 3). A distribution of the RGB pixel values in each square was made, allowing the mean pixel and standard deviation values to be determined for each 32 $\times$ 32 pixel square. A square was considered to be important if the mean pixel value ($\pm$ standard deviation for a 32 $\times$ 32 block) was greater or less than either the mean pixel value of the specific frame ($\pm$ pixel standard deviation for a specific frame) or the mean pixel value for the specific video clip ($\pm$ pixel standard deviation for the entire video). As considerable variation in color, level of edges, and movement is possible both in a particular frame and throughout the video, it was considered equally important to include both conditions.

The 16-pixel shift between squares ensures that the majority of the image is covered by four separate comparisons. However, the image edges are only covered by two comparisons, and the corner 16 $\times$ 16 pixel squares are only covered by one result. Areas deemed as having important content are written to an output file, called a RoI script, which contains all RoI squares for each frame of each multimedia video clip. This script, as in the case of eye-tracking data, is subsequently used to adapt RoIDs. Content-dependent RoI for a specific frame can be seen in Fig. 4.

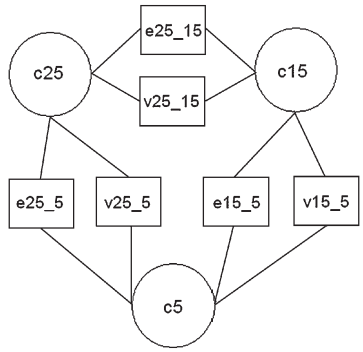

Fig. 5. Shows the nine video combinations (c = control; e = empirically defined video; v = predicted defined video). The first number represents the foreground frame rate, and the second number represents the background frame rate. Control video implements multiple-frame-rate video.

3) Video Creation Using Empirical and Predictive RoID Data: To create RoIDs, we are required to produce video that has an adaptive nonuniform distribution of resource allocation. To achieve this, we used empirical eye-based and predictive content-dependent RoI data to adapt the frame rate in particular regions of the screen. Thus, it was decided that RoI areas, herewith referred to as foreground areas, should be refreshed at a relatively higher frame rate than that of the non-RoI areas (background areas).

Software was developed, using the Java Media Framework, which took the original video (at 25 fps) and a RoI script (either containing empirical or predictive RoI data) and, using a five-frame buffer, produced a playable RoID, which presents the foreground (RoI) and background (non-RoI) regions at different frame-rate combinations (see Fig. 6). At playback, this video can be considered as a RoID, as it plays defined RoI at a higher "quality" than that of peripheral areas.

To identify how varied foreground and background frame-rate combinations impact user perception, our study considers three quality combinations for the control videos, and empirical and predictive RoIDs. Accordingly, nine potential *(display, frame rate)* combinations were considered as part of our perceptual experiments: control video (no background) at 25 fps (**c25**), control video (no background) at 15 fps (**c15**), control video (no background) at 5 fps (**c5**), empirically defined RoI with 25 fps in the foreground and15 fps in the background (**e25_15**), empirically defined RoI with 25 fps in the foreground and 5 fps in the background (**e25_5**), empirically defined RoI with 15 fps in the foreground and 5 fps in the background (**e15_5**), predicted RoI with 25 fps in the foreground and 15 fps in the background (**v25_15**), predicted RoI with 25 fps in the foreground and 5 fps in the background (**v25_5**), and predicted RoI with 15 fps in the foreground and 5 fps in the background (**v15_5**). This is shown in Figs. 5 and 6.

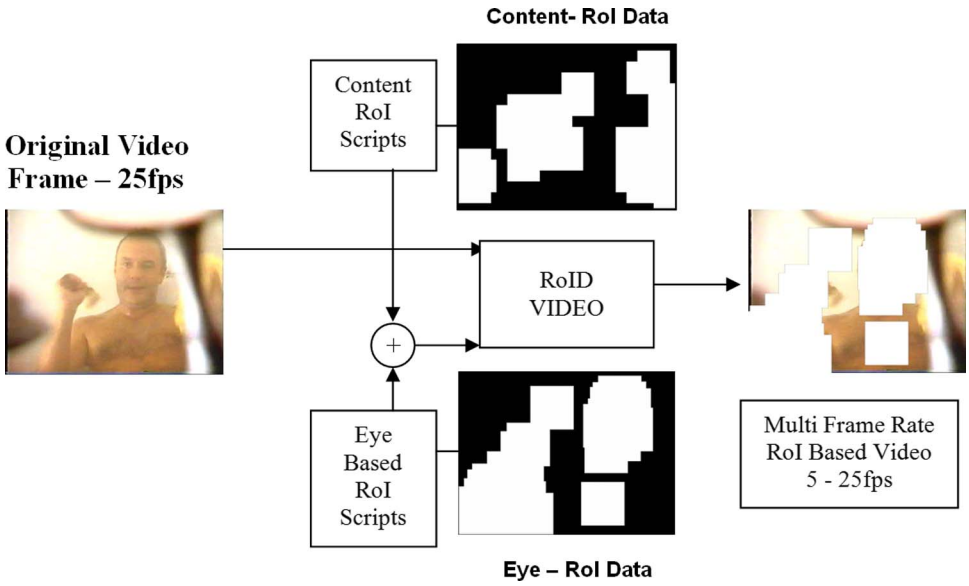

Fig. 6. Process of adapting RoIDs from eye-based content-dependent data. White area in output video signifies higher refresh area.

D. Phase 3: Perceptual Experiments

Perceptual experiments were carried out to answer the following questions.

1) *Does the use of eye-based (empirical) or content-dependent (predictive) RoIDs impact a user's level of IA and understanding?*
2) *Does the use of empirical or predictive RoIDs impact a user's perceived subjective LoQ or a user's LoE?*

1) Participants: In our perceptual experiment, we wanted participants to view all nine potential *(display, frame rate)* combinations—to ensure that all RoID display types and frame-rate combinations were viewed by each participant (see Fig. 5). Accordingly, in order to ensure that all 12 video clips were shown at all qualities, we required nine experimental groups. Fifty-four participants—different from those in phase 1—were evenly divided into nine experimental groups. Participants in this phase of our study were aged between 21 and 67. Analogously to phase 1 of the study, participants volunteered to take part, were drawn from the authors' contacts, and represented a mix of different nationalities and backgrounds.

2) Perceptual Experiment—Setup: To ensure that experimental conditions remained consistent, the same experimental equipment was used for all participants. An HP mobile laptop AMD Athlon XP 2000+, with an inbuilt 15-in LCD monitor and a ATI Radeon IGP 320M, was used to display video with a resolution of 640 × 480. Video clips, as described in Section IV-D, were embedded in an Internet Explorer browser, thus simulating realistic conditions under which RoIDs might be used. To ensure that a consistent audio level (70 dB) was used for all participants, headphones were used when a video was playing. Once the laptop and headphones were appropriately set up and the user felt comfortable with the position of the screen, the experimental process, defined in Section IV-B3, was applied. Instead of closing the video window, a blank webpage was used to hide the video after each viewing.

V. Results

The aim of attentive displays is to provide a high level of perceived quality, with the least use of processing, memory, and bandwidth resources. To achieve this, we need to better understand the two dimensions of infotainment, both information transfer and user satisfaction. Accordingly, the following results consider the impact that RoID manipulation quality (see Fig. 5) has on user IA (QoP-IA), user perception of quality (QoP-LoQ), and user perception of enjoyment (QoP-LoE).

A. Impact of RoID Quality and Clip Type on QoP-IA

QoP-IA (a user's level of IA/understanding) was expressed as a percentage measure, which reflected the level of information assimilated from visualized multimedia content. A multiple analysis of variance test was used with *display type (control, empirical, predictive)* and *minimum video frame rate* (5 fps: c5, e25_5, e15_5, v25_5, v15_5; 15 fps: c15, e25_15, v25_15; or 25 fps: c25) as separate independent variables. This revealed that combined display and minimum frame-rate combinations do not have a significant effect on user QoP-IA $\{F(1,2) = 0.223, p = 0.800\}$ [see Fig. 7(a)]. When analyzed separately, no significant QoP-IA variation was identified as a result of the display type (control, empirical, or predictive RoIDs) or due to foreground/background combinations (i.e., 25/15, 25/5, or 15/5 fps). This finding complements a previous work targeting non-RoI video [9] and suggests that adaptation of video playback can take place without detrimentally impacting user understanding of the video content. This result implies that users are still able to absorb the critical information despite considerable frame loss and perceptual issues introduced by RoID. Therefore, we can claim that video frame loss does not prevent the user from understanding the video narrative.

Interestingly, a multiple analysis of variance test, with *display*, *minimum frame rate*, and *video clip* as separate fixed factors, showed that a user's ability to assimilate and understand information from certain video clips was significantly impacted

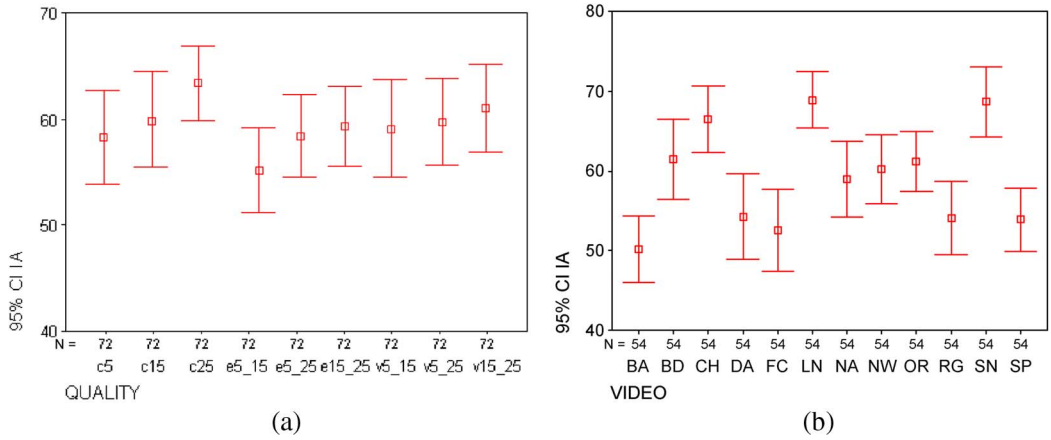

Fig. 7. Mean and 95% significance plot for QoP-IA—dependent on (a) quality and (b) video type.

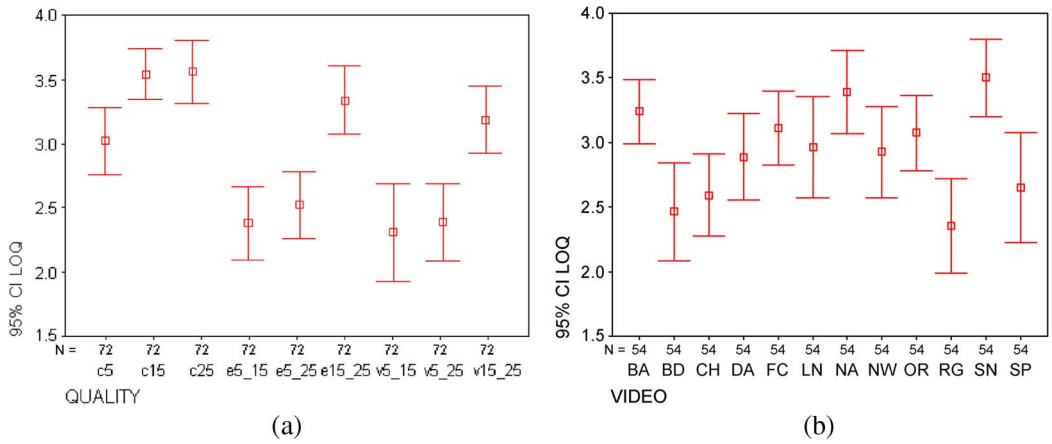

Fig. 8. Mean and 95% CI plot for QoP-LoQ, dependent on (a) display and frame-rate combinations and (b) video type.

$\{F(1,22) = 1.879, p = 0.09\}$. This suggests a significant variation in the level of information being assimilated when watching specific video clips. It is noteworthy to observe the significant variation in user IA $\{F(1,11) = 6.771, p < 0.001\}$ that exists as a result of the video clip type [see Fig. 7(b)]. Results show that, despite a constant number of questions, the type of video being presented is more significant to the level of user IA than the quality at which the video is being presented.

This result is of considerable interest, particularly in the fields of advertising and education, as it implies that in order to ensure that users understand and assimilate the optimum level of information, the type of video being presented is more significant to user QoP-IA than the quality of the video presentation. Beyond the scope of this paper, additional research is needed to more fully investigate relationships that exist between video type and QoP-IA.

B. Impact of RoID Quality and Clip Type on QoP-LoQ

User LoQ (QoP-LoQ) defines the user's subjective opinion concerning video QoS. A multiple analysis of variance test showed QoP-LoQ as being independently affected by *display type* $\{F(1,2) = 9.071, p < 0.001\}$, *minimum frame rate* $\{F(1,2) = 38.196, p < 0.001\}$ [see Fig. 8(a)], and type of video $\{F(1,11) = 5.831, p < 0.001\}$ [see Fig. 8(b)], as well as the combined effect of all multiple factors $\{F(1,22) = 2.166, p = 0.002\}$. QoP-LoQ was significantly different when videos were presented using different techniques (i.e., control versus RoID). However, no variation occurred as a result of using specifically empirical or predictive RoID techniques (Post-Hoc Tukey-Test: $p = 0.632$). Results imply that the user is aware of certain quality reduction in video, even when this reduction does not impact upon the user's overall level of understanding (QoP-IA). Accordingly, just because the user understands the content of the video does not mean s/he accepts the video as being of good quality. The risk, particularly when manipulating educational material, is to assume that information transfer is sufficient. These results, however, imply that perception of quality and general satisfaction should be measured and manipulated as separate factors. Moreover, these results state that the video content is essential to user perception of quality, and imply that this should be considered in addition to traditional technical quality criteria.

Post-Hoc Tukey-Tests showed a significant difference in QoP-LoQ between control videos shown at 5 fps, compared to those shown at both 15 fps $\{p = 0.040\}$ and 25 fps $\{p = 0.031\}$. No significant difference was measured between QoP-LoQ when participants were shown control videos at 15 and 25 fps, suggesting that participants view 15 and 25 fps as being of similar quality [this is shown in Fig. 8(a)]. Results revealed significant differences in the level of QoP-LoQ between videos shown with a foreground/background

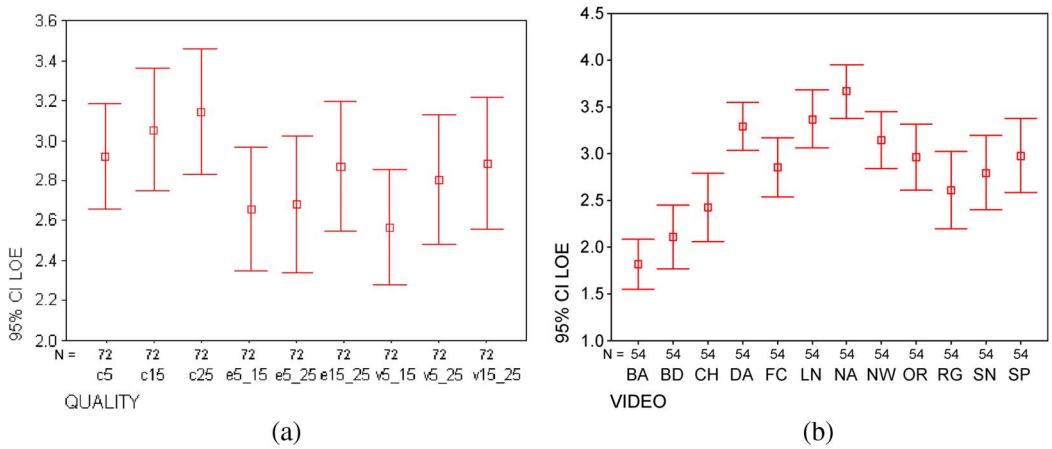

Fig. 9. Mean and 95% significance plot for QoP-LoE, dependent on (a) quality and (b) video type.

combination of 25/15 fps and all other RoID videos, independent of the display approach $\{e25/15 - e25/5 : p < 0.001;\ e25/15 - e15/5 : p < 0.001;\ \mathrm{v}25/15 - \mathrm{v}25/5 : p < 0.001;\ \mathrm{v}25/15 - \mathrm{v}15/5 : p < 0.001\}$. This implies that the use of multi-frame-rate adapted RoIDs, with a background frame rate of less than 15 fps, will negatively affect the user perception of quality. This supports the work of Wijesekera *et al.* [36] that suggests that frame rate should be maintained at or above 12 fps if the user's perception of quality is to be maintained. Accordingly, all video-based manipulation should endeavor to maintain a frame rate of at least 15 fps if user perception of quality is to be maintained.

C. Impact of RoID Quality and Clip Type on QoP-LoE

QoP-LoE is the subjective LoE experienced by a user when watching a multimedia presentation. It is intuitive to assume that, as a result of personal preference, the type of video being presented to a participant will significantly affect a user's LoE. This is supported by our work $\{F(1, 11) = 8.911, p < 0.001\}$ and can be clearly seen in Fig. 9(b). Further work is required to identify whether a relationship exists between LoE, variations in user demographic, video clip content, and presentation style. Such research, in support of findings in Section V-B, would help increase user enjoyment, which, in turn, could be used to augment user quality perception.

More interestingly, our results identify that user LoE is not significantly affected by combined *display* and *minimum frame-rate* combinations $\{F(1, 2) = 0.16, p < 0.985\}$ [see Fig. 9(a)], even though we have already identified earlier that users are able to effectively distinguish between the quality of video presentations. Accordingly, it seems that independent of whether the video was perceived as being enjoyable, users can distinguish between the video quality and their subjective LoE. This result shows that the LoE is more significantly related to the type of video being presented than the technical quality of the presentation. This is an interesting result, particularly in the field of entertainment and/or within bandwidth-constrained environments, as it suggests that changes in the presentation approach do not significantly impact a user's LoE. Moreover, it supports the claim that effective content adaptation and personalization of information could be used to improve user perception where technical limitations are identified.

VI. Conclusion

The user perspective of multimedia video quality, although an important determinant of quality, is rarely used to adapt media content in current distributed multimedia environments. In this paper, we have presented the results of a study, which used empirically captured eye-tracking data and predicted content primitive data to allow us to implement two RoID techniques. In our study, perception encapsulated multimedia's infotainment duality and was understood as not only a user's understanding of the video content (QoP-IA) but also his/her subjective opinion concerning video: regarding QoS provision (QoP-LoQ) and perceived LoE (QoP-LoE).

Results show that the presentation method (i.e., control video, empirical RoID, and predictive RoID) does not significantly impact user ability to understand video narrative. This result highlights the fact that the expensive task of preliminary gathering of eye-tracking data, to define RoI, is not necessary if effective automatic processing can be achieved. If automatic processing allows effective frame-based adaptation of video, then real-time automatic RoID players could be developed. Conversely, a considerable mean variation in user understanding was measured as a result of the video clip type; moreover, both types of adaptation (empirical and predictive) considered in this paper, as well as the type of video being presented, affect user perceived quality. This means that RoID frame-rate displays are unable to maintain a high perceived LoQ, if frame rates less than 15 fps are used. This has interesting implications, both for producers of media content as well as for its transmission. Further research is required to investigate better ways of extracting, mapping, and displaying variable quality video content.

Our findings are of particular interest for the transmission of entertainment-oriented multimedia in bandwidth-constrained/-hungry environments [2], [26], since they show that while user LoE was significantly affected by the type of video being shown, QoP-LoE was not affected by *display* and *minimum frame-rate* combinations. Nonetheless, the implications of our work go beyond entertainment-related content, because they could be potentially applied to any application area associated with transmission of video in bandwidth-limited environments or video storage on systems in which computer memory is at a premium—real-time traffic monitoring and surveillance, tele-surgery and electronic storage of video-intensive medical

records, immersive vision applications, e-learning, to mention a few. All represent avenues for future research, and our work highlights the potential advantages of using attentive RoIDs in these contexts.

REFERENCES

[1] P. Baudisch, D. DeCarlo, A. T. Duchowski, and W. S. Geisler, "Focusing on the essential: Considering attention in display design," *Commun. ACM*, vol. 46, no. 3, pp. 60–66, Mar. 2003.

[2] L. Chen and B. Veeravalli, "Multiple-server movie-retrieval strategies for distributed multimedia applications: A play-while-retrieve approach," *IEEE Trans. Syst., Man, Cybern. A, Syst., Humans*, vol. 36, no. 4, pp. 786–803, Jul. 2006.

[3] B. L. Cole and P. K. Hughes, "Drivers don't search: They notice," in *Visual Search*, D. Brogan, Ed. New York: Taylor & Francis, 1990, pp. 407–417.

[4] M. Czerwinski, D. S. Tan, and G. G Robertson, "Women take wider view," in *Proc. CHI*, Minneapolis, MN, Apr. 2002, pp. 195–202.

[5] A. T. Duchowski and A. Coltekin, "Foveated gaze-contingent display for peripheral LOD management, 3D visualisation, and stereo imaging," *ACM Trans. Multimedia Comput., Commun. Appl.*, vol. 3, no. 4, pp. 1–18, 2007.

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

[7] J. M. Findlay, "The visual stimulus for saccadic eye movements in human observers," *Perception*, vol. 9, no. 1, pp. 7–21, 1980.

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

[9] G. Ghinea and J. P. Thomas, "QoS impact on user perception and understanding of multimedia video clips," in *Proc. ACM Multimedia*, Bristol, U.K., 1998, pp. 49–54.

[10] G. Ghinea, "Quality of perception—An essential facet of multimedia communications," Ph.D. dissertation, Dept. Comput. Sci., Univ. Reading, U.K., 2000.

[11] S. R. Gulliver and G. Ghinea, "Stars in their eyes: What eye-tracking reveals about multimedia perceptual quality," *IEEE Trans. Syst., Man, Cybern. A, Syst., Humans*, vol. 34, no. 4, pp. 472–482, Jul. 2004.

[12] A. P. Hillstrom and S. Yantis, "Visual motion and attentional capture," *Percept. Psychophys.*, vol. 55, no. 4, pp. 399–411, Apr. 1994.

[13] D. H. Hubel and T. N. Wiesel, "The period of susceptibility to the physiological effects of unilateral eye closure in kittens," *J. Physiol.*, vol. 206, no. 2, pp. 419–436, Feb. 1970.

[14] M. A. Just and P. A. Carpenter, "Eye fixations and cognitive processes," *Cogn. Psychol.*, vol. 8, pp. 441–480, 1976.

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

[16] L. Kaufman and W. Richards, "Spontaneous fixation tendencies for visual forms," *Percept. Psychophys.*, vol. 5, no. 2, pp. 85–88, 1969.

[17] N. Krüger, M. Lappe, and F. Wörgötter, "Biologically motivated multimodal processing of visual primitives," *Interdiscip. J. Artif. Intell. Simul. Behav.*, vol. 1, no. 5, pp. 417–428, 2004.

[18] N. Krüger, M. Felsberg, and F. Wörgötter, "Processing multi-modal primitives from image sequences," in *Proc. 4th Int. ICSE Symp. EIS*, Madeira, Portugal, 2004.

[19] L. C. Loschky and G. W. McConkie, "User performance with gaze-contingent multiresolutional displays," in *Proc. ACM ETRA*, Palm Beach Gardens, FL, 2000, pp. 97–103.

[20] L. C. Loschky and G. S. Wolverton, "How late can you update gaze-contingent multi-resolutional displays without detection?" *ACM Trans. Multimedia Comput., Commun. Appl.*, vol. 3, no. 4, pp. 1–10, 2007.

[21] J. F. Mackworth and J. S. Bruner, "How adults and children search and recognize pictures," *Hum. Dev.*, vol. 13, no. 3, pp. 149–177, 1970.

[22] A. Mahanti, D. L. Eager, M. K. Vernon, and D. J. Sundaram-Stukel, "Scalable on-demand media streaming with packet loss recovery," *IEEE/ACM Trans. Netw.*, vol. 11, no. 2, pp. 195–209, Apr. 2003.

[23] J. T. Mordkoff and S. Yantis, "Dividing attention between color and shape: Evidence of coactivation," *Percept. Psychophys.*, vol. 53, no. 4, pp. 357–366, Apr. 1993.

[24] W. Osberger and A. J. Maeder, "Automatic identification of perceptually important regions in an image using a model of the human visual system," in *Proc. 14th Int. Conf. Pattern Recog.*, Brisbane, Australia, 1998, vol. 1, pp. 701–704.

[25] W. Osberger, A. J. Maeder, and N. Bergmann, "A perceptually based quantization technique for MPEG encoding," in *Proc. SPIE Human Vis. Electron. Imaging III*, Jan. 26–29, 1998, vol. 3299, pp. 148–159.

[26] S. Polak, Y. Barniv, and Y. Baram, "Head motion anticipation for virtual-environment applications using kinematics and EMG energy," *IEEE Trans. Syst., Man, Cybern. A, Syst., Humans*, vol. 36, no. 3, pp. 569–576, May 2006.

[27] E. M. Reingold and L. C. Loschky, "Reduced saliency of peripheral target in gaze-contingent multi-resolutional displays: Blended versus sharp boundary windows," in *Proc. ACM ETRA*, New Orleans, LA, 2002, pp. 89–93.

[28] E. M. Reingold, L. C. Loschky, D. M. Stampe, and J. Shen, "An assessment of a live-video gaze contingent variable resolution display," in *Proc. Human-Comput. Interaction*, M. J. Smith, G. Salvendy, D. Harris, and R. J. Koubek, Eds., 2001, pp. 1338–1342.

[29] S. Saida and M. Ikeda, "Useful visual field size for pattern perception," *Percept. Psychophys.*, vol. 25, no. 2, pp. 119–125, Feb. 1979.

[30] A. F. Sanders, "Some aspects of the selective process in the functional visual field," *Ergonomics*, vol. 13, no. 1, pp. 101–117, Jan. 1970.

[31] S. Shioiri and M. Ikeda, "Useful resolution for picture perception as a function of eccentricity," *Perception*, vol. 18, no. 3, pp. 347–361, 1989.

[32] P. L. Silsbee, A. C. Bovik, and D. Chen, "Visual pattern image sequence coding," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 3, no. 4, pp. 291–301, Aug. 1993.

[33] L. B. Stelmach, W. J. Tam, and J. Hearty, "Static and dynamic spatial resolution in image coding: An investigation of eye-movements," in *SPIE Human Vis., Vis. Process. Digital Display*, 1992, vol. 1453, pp. 147–152.

[34] A. Tresman, "Features and objects in visual processing," *Sci. Amer.*, vol. 255, no. 5, pp. 106–115, Nov. 1986.

[35] J. A. Turner, "Evaluation of an eye-slaved area-of interest display for tactical combat simulation," in *Proc. 6th Interservice/Ind. Training Equipment Conf. Exhib.*, 1984, pp. 75–86.

[36] D. Wijesekera, J. Srivastava, A. Nerode, and M. Foresti, "Experimental evaluation of loss perception in continuous media," *Multimedia Syst.*, vol. 7, no. 6, pp. 486–499, Nov. 1999.

[37] A. L. Yarbus, *Eye Movement and Vision*. New York: Plenum, 1967. (trans. B. Haigh).

Stephen R. Gulliver (M'02) received the B.Eng. (Hons.) degree in microelectronics, the M.Sc. degree in distributed information systems, and the Ph.D. degree from Brunel University, Uxbridge, U.K., in 1999, 2001, and 2004 respectively.

He is currently a Lecturer with the Informatics Research Centre, University of Reading, Reading, U.K. His research interests focus on pervasive informatics and includes human factors, 3-D and virtual reality accessibility, attention analysis, and the perceptual and information acquisition aspects of computer and multimedia systems.

Gheorghita Ghinea (M'02) received the B.Sc. and B.Sc. (Hons.) degrees in computer science and mathematics, the M.Sc. degree in computer science from the University of the Witwatersrand, Johannesburg, South Africa, in 1993, 1994, and 1996, respectively, and the Ph.D. degree in computer science from the University of Reading, Berkshire, U.K., in 2000.

He is currently a Reader with the Department of Information Systems and Computing, School of Information Systems, Computing and Mathematics, Brunel University, Uxbridge, U.K. His research interests span perpetual aspects of multimedia, quality of service, multimedia resource allocation, and computer networking and security issues.