Correspondence

Quality of Perception: User Quality of Service in Multimedia Presentations

Gheorghita Ghinea, *Member, IEEE,* and Johnson P. Thomas

***Abstract*—We define quality of perception (QoP) as representing the user side of the more technical and traditional quality of service. QoP encompasses not only a user's satisfaction with the quality of multimedia presentations, but also his/her ability to analyze, synthesise and assimilate the informational content of multimedia displays. We found that significant reductions in frame rate and color depth does not result in a significant QoP degradation.**

***Index Terms*—Frame rate, multimedia video, quality of perception.**

I. Introduction

Although the effectiveness of multimedia applications depends largely on the performance capabilities of networking protocols and communication delivery systems, optimum service, however, cannot always be guaranteed due to two competing factors: multimedia data sizes and network bandwidth. When these factors degrade a network's performance, congestion, packet loss, bit errors and out-of-order arrivals result. Consequently, a great deal of research in this area has focused on the technical and networking aspects of delivering multimedia applications. The success of a particular application, however, is ultimately determined by the end-user's experience.

Research into the end-user's perception of and satisfaction with multimedia applications delivered over networks has been relatively limited, however. In this context, Apteker *et al.* have investigated the effects that different video frame rates have on human satisfaction with the multimedia presentation [1]. Their results showed that for certain ranges of human receptivity, a small variation of it leads to a much larger relative variation of the required bandwidth. Closely related to this work is the one of Fukuda *et al.* who derived a mapping between the required bandwidth of multimedia video and three quality of service (QoS) parameters (frame rate, signal-to-noise ratio, spatial resolution) [4], whilst Yamazaki examined the effects of different frame rates, sizes and quantization parameters of MPEG-4 video on perceptual quality [8].

Blakowski and Steinmetz showed that synchronization between media is generally characterized by three regions: one in which synchronization errors are unnoticeable by the user, one in which they are perceived but tolerated, and one in which they are found irritating [2]. Kawalek, on the other hand, is more interested in the cut-off rate beyond which the quality of transmitted audio and video becomes unacceptable to human users in desktop conferencing environments [6]. He showed that the perception of media loss is highly dependent on the medium in question. While Bouch *et al.* have researched the effect of latency on perceived Web QoS [3], Wijesekera *et al.* build on Steinmetz's and Apteker's earlier work and investigate the perceptual tolerance to discontinuity caused by media losses and repetitions, and to that of varying degrees of missynchronization across streams [9].

User satisfaction, perception and understanding of multimedia should be the driving force in networking and operating systems research. The focus of our work has been the enhancement of the traditional view of QoS with a user-level defined *quality of perception* (QoP). This measure encompasses not only a user's satisfaction with multimedia clips (which we shall denote by $\mathrm{QoP_S}$), but also his/her ability to understand, synthesise and analyze the informational content of such presentations (which we shall denote by $\mathrm{QoP_U}$). We believe that a measure such as QoP will have more meaning for a typical multimedia user than typical QoS metrics. As such, we have investigated the interaction between QoP and QoS and its implications from both a user perspective as well as from a networking angle.

II. Experiments

Our approach to evaluating QoP has been mainly empirical, as is dictated by the fact that its primary focus is on the human-side of multimedia computing. Each user was presented with 12 windowed ($352 * 288$ pixels) MPEG-1 video clips. Each of the clips was between 31 and 45 s long. Subjects were randomly selected from a volunteer pool comprising persons of various backgrounds and professions (school children, students, educators, white-collar workers, as well as clerical and administrative staff). All were computer literate. In order to measure $\mathrm{QoP_U}$, after the users had seen each clip once, the window was closed, and they were asked between 10–12 questions (depending on the video clip) about its informational content. Users were then asked to rate $\mathrm{QoP_S}$ on a scale of 1 – 6 (with scores of 1 and 6 representing the worst and, respectively, best perceived qualities possible). The user then visualised the next clip.

Each clip was shown with the same set of QoS parameters, unknown to the user. In our work we have focused on application-level QoS parameters, as we were especially interested in the perceptual impact that variations in their value entail, and how these results could be explored to make more efficient use of network bandwidth, the QoS parameter characterising what is arguably the most scarce resource in distributed multimedia systems.

In our experiments, QoS parameters were not modified in the case of the audio stream. The reason why it was not decided to operate on the audio stream was primarily because humans are more susceptible to perceive audio loss, rather than video [6]. Even in the case of low frame rate multimedia video sequences, the compressed audio stream requires less bandwidth than the video. It thus makes more sense to operate on the video component, where there is more scope to achieve bandwidth gains without significant perceptual loss. This choice is further strengthened by the fact that, due to its scarcity, bandwidth is the main resource we are ultimately interested in using more efficiently. Parameters were thus varied in the case of the video stream. These include both spatial parameters (color depth) and temporal parameters (frame rate). Two different color depths were used (8 and 24-bit), together with three different frame rates (5, 15 and 25 frames per second – fps). A total of 72 users, 12 for each (*frame_rate*, *color_depth*) parameter pair, were tested.

Even though (*frame_rate*, *color_depth*) parameters were varied across the experiments, for a particular user they were kept constant. Users were furthermore kept unaware of the values of the parameters

Manuscript received April 15, 2002; revised January 28, 2004. The associate editor coordinating the review of this manuscript and approving it for publication was Dr. Minoru Etoh.

G. Ghinea is with the Department of Information Systems and Computing, Brunel University, Middlesex UB8 3PH, U.K. (e-mail: George.Ghinea@brunel.ac.uk).

J. P. Thomas is with the Computer Science Department, Oklahoma State University, Tulsa, OK 74106 USA (e-mail: jpt@okstate.edu).

Digital Object Identifier 10.1109/TMM.2005.850960

Fig. 1. Snapshots of the multimedia clips used in QoP experiments (from left to right: *pop music, cooking, rugby, documentary*).

TABLE I
VIDEO CATEGORIES

VIDEO CATEGORY	Dynamic	Audio	Video	Text
1 - Action Movie				
2 - Animated Movie				
3 – Band				
4 – Chorus				
5 – Commercial				
6 – Cooking				
7 – Documentary				
8 – News				
9 - Pop Music				
10 – Rugby				
11 – Snooker				
12 - Weather Forecast				

with which the clips were being shown to them. A between-subjects design was chosen for our experiments as it was unreasonable to show subjects the same video clip transmitted with different QoS parameters and then ask them identical questions regarding the subject matter. If this would have been done, obviously the number of questions answered correctly would have been much higher the second time round they saw the clip and the results obtained would be of little value.

In order to eliminate any possible individual differences (e.g., better memory, greater keenness or interest in the subject matter) in the way the different subjects handled the experimental task, subjects were allocated randomly to each parameter pair. Moreover, special care was taken for the results not to be influenced by any *order effects*, whereby a presentation of the video clips in the same order might facilitate users scoring low marks in the initial sequences when they are getting used to the test environment and methodology. Thus, in order to counterbalance any order effects, the order of presentation of the clips was varied from user to user, on a cyclic basis.

The clips themselves were digitized in MPEG-1 format and chosen to cover a broad spectrum of infotainment subject matter, ranging from a relatively static news clip to a highly dynamic rugby football sequence (Fig. 1). All depicted excerpts from real-world programmes and thus represent informational sources which an average user might encounter in everyday life. Table I contains a classification of the clips taking into account temporal parameters (how dynamic/static the clips were), as well as the importance of audio, video and text as conveyors of information in the context of the clip. A dark grey table cell indicates a highly dynamic video clip or a clip where the respective medium (audio, video or text) is highly important in carrying information. A light-grey cell indicates that a clip is of medium dynamicity or the medium is of medium importance. A white table cell signifies a static clip or the medium is of little or no importance. This classification was obtained by asking 12 respondents (other than the ones selected later for the experiments proper) to rate each clip (run at 25 fps and 24-bit color depth) on a 7-point Likert scale.

For each clip, the $\mathrm{QoP_U}$ questions were chosen to encompass all aspects of the information – audio, visual or textual – presented in the clips. Additionally, some questions could only have been answered if the user had grasped pieces of both visual and audio information from the clip. Lastly, although there were no "trick" questions as such, quite a few of them could not be answered by observation of the video alone, but by the user making inferences and deductions from the information that had just been presented.

III. Analysis of Results

A. $\mathrm{QoP_U}$

One of the important findings of this work is that if the primary purpose of the multimedia exercise is educational (i.e., the assimilation of information), then this component of Quality of Perception does not vary with vastly reduced bandwidth requirements. A three-way analysis of variance (*ANOVA*) was performed on the experimental data with frame rate, color depth and clip type as independent variables. The analysis showed that $\mathrm{QoP_U}$ does not significantly vary with color depth or frame rate, and neither with the interaction between any of the three variables considered (Figs. 2 and 3). This finding has important implications in bandwidth-constrained environments, for it highlights that the $\mathrm{QoP_U}$ for a multimedia clip shown at 25 fps with 24-bit color depth does not change significantly if severe bandwidth limitations force it to be shown at 5 fps with 8-bit color. Indeed, in some cases, user $\mathrm{QoP_U}$ was marginally higher at lower frame rates. This could be explained by the fact that the complementary process to frame dropping is one of frame replication. Due to the latter, information that might had been lost had the clip been played with its designated frame rate, would now appear for a longer period of time (three or even five times longer in the case of our experiments) on the screen. This would therefore increase the chance of the user noticing the respective information.

We were also interested in determining if the type of clip had an effect on $\mathrm{QoP_U}$. The above 3-way ANOVA of the $\mathrm{QoP_U}$ results showed that the percentage of correct answers obtained were very different for different clips, with a level of significance of 0.05 ($\mathrm{F}(11, 66) = 2.07$, $\mathrm{p} < .05$ where 11 and 66 are degrees of freedom and 0.05 is the confidence level). We took our analysis further by doing a least significant differences (*LSD*) test to isolate groups of clips which are different from each other. Within each derived subset, the highest and lowest means of the percentage of correct answers will not be significantly different.

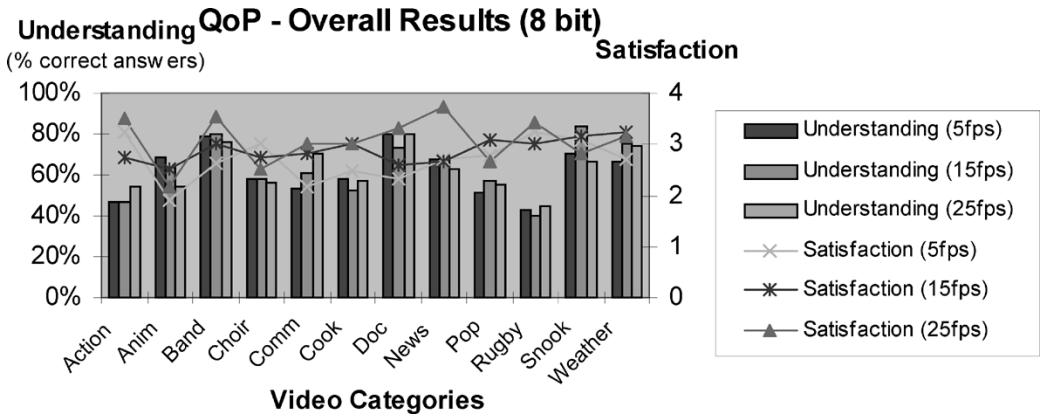

Fig. 2. Experimental QoP results at 8-bit color depth.

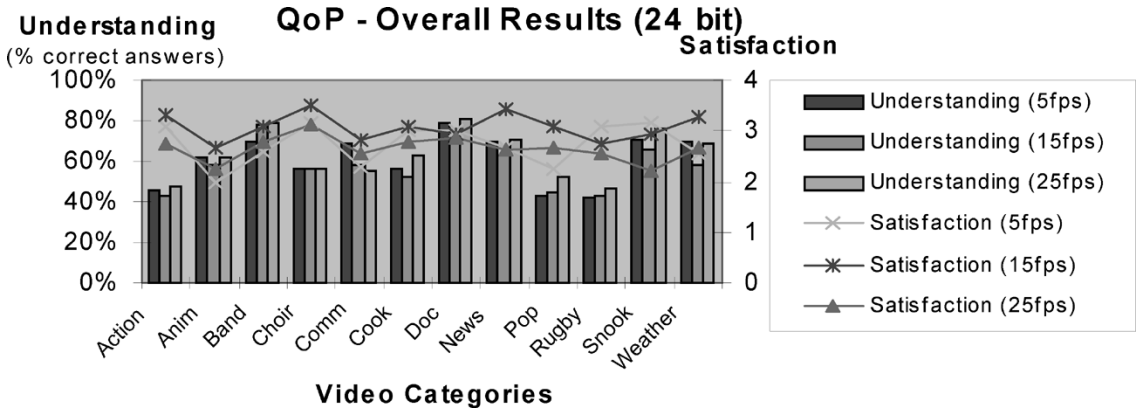

Fig. 3. Experimental QoP results at 24-bit color depth.

If the significance level is taken to be 0.05, the following subsets are isolated:

$$A = (Action, Rugby); B = (Rugby, Pop);$$
$$C = (Cooking, Chorus, Commercial, Animation)$$
$$D = (Animation, News); E = (News, Weather, Snooker);$$
$$F = (Snooker, Band, Documentary).$$

This analysis shows that there is a strong link between $\mathrm{QoP_U}$ and clip content. Thus, for example, the two clips comprising subset B have a fairly close content (see Table I) and similar $\mathrm{QoP_U}$. A few further interesting observations can be made. The first is that users tend to have a low $\mathrm{QoP_U}$ when highly dynamic clips, with rapid scene changes are involved (as is the case of subset A, comprising the Action and Rugby clips). When clip scenes are varying rapidly, it is of course difficult to get any sort of visual information, the most one can do is abstract the message of the clip. The fact that frame dropping has little impact here should not, therefore, surprise.

A low $\mathrm{QoP_U}$ is also scored in the case of clips which are moderately dynamic, but are very rich in information and use all the available media to transmit this information, as is the case of the Pop music clip, contained in subset B. In the case of such clips, what usually happens is that users cannot distribute their attention, which explains why respondents got such low percentages of correct answers. Lastly, users' $\mathrm{QoP_U}$ had median values for the largest subset isolated, subset C, all of whose clips involve 'talking heads'. This is in contrast to the Documentary (static clip, with highest $\mathrm{QoP_U}$) and Action clips (highly dynamic clip, with lowest $\mathrm{QoP_U}$), where the audio is narrated by an unseen speaker. The suggestion therefore is that in the case of clips with "talking heads" users focus a lot on the speaker, thus ensuring that although $\mathrm{QoP_U}$ will never be at a minimum (speakers are an important source of information), it will never peak either (since concentrating attention on one informational source leaves other potential sources unnoticed). This conclusion is reinforced by the fact that all the clips in subset F (best $\mathrm{QoP_U}$ ratings) do not involve "talking heads."

A 1-way ANOVA was further performed to establish whether the relative importance of the video, audio and textual components within the context of a clip (as given in Table I) has any impact on user $\mathrm{QoP_U}$. In this case, at a significance level of 0.05 ($\mathrm{F}(11,122) = 1.92, \mathrm{p} < .05$), we found that the correct answers given by respondents was dependent upon the information load conveyed by the video, audio and textual contents of the multimedia clips. Extending this study with a LSD test with significance level 0.05, we get that variations in the observed QoP are mainly due to differences between the informational content of text and video. We obtain the following ordering in terms of correct answers obtained for the different video clips: $Text > Audio > Video$

Video and text are both visual media and it appears that a user cannot concentrate on both. Users seem to be able to focus on the audio and one of the visual media at a time, a finding in accordance with previous work [7] (which, however, assumed optimal multimedia presentation quality). The fact that the highest number of correct answers was given when information was conveyed textually can be explained by the fact that the textual component of the multimedia clips used in our experiments was static and (in the majority of cases) conspicuous. It thereby easily attracts the visual attention of the subject, which explains its leading position as conveyor of information. This also explains why, although the audio stream was transmitted at full quality in the experiments, textual information was more easily retained than the audio.

Thus, users seem to prioritise informational sources. This is also confirmed by the finding that information contained in the video stream fared the worst as far as $\mathrm{QoP_U}$ goes: as video was often shown with low quality, users probably tended to concentrate more on the audio stream and disregard video impediments.

B. $\mathrm{QoP_S}$

We now consider the second component of QoP, namely $\mathrm{QoP_S}$, the user's satisfaction with the presentation. A 3-way ANOVA was done with frame rate, color depth and clip type as independent variables and showed that, although $\mathrm{QoP_S}$ does not significantly depend on the frame rate or color depth at which clips were shown, it does depend strongly on the interaction between frame rate and color depth, with a level of significance of 0.01 ($\mathrm{F}(2, 66) = 5.23, \mathrm{p} < .01$). Thus, if both parameters are simultaneously changed within a single presentation, user satisfaction is very likely to be affected (Figs. 2 and 3).

Previous experiments [1], [4] studying the effect of frame rate variation on users' satisfaction have shown that at low frame rates there is a dramatic improvement in the satisfaction with the perceived quality of the video as the frame rate increases. However, at high frame rates there is a minimal variation in the user's satisfaction, with increasing frame rates, a characteristic also called *asymptotic behavior*. In contrast, our results would seem to indicate that, when told that they would actually be examined on the informational content of the clips, users concentrate on absorbing the information present in the clips and little appear to notice the QoS degradation in the clips sent over the network.

The three-way ANOVA done above also highlighted that clip type has a statistically significant impact, with a significance level of 0.01 ($\mathrm{F}(11, 66) = 2.98, \mathrm{p} < .01$), on $\mathrm{QoP_S}$. This analysis was extended by performing an LSD test with significance level of 0.05. The test isolated three major subsets which, in order of decreasing user satisfaction, are

$$A = (Animation, Commercial)$$

$$B = (Commercial, Pop, Documentary, Cooking, Snooker, Weather, Band, News)$$

$$C = (Chorus, Rugby, Action).$$

What can be observed from this classification is that although dynamic clips, such as Rugby and Action, score very low as far as both $\mathrm{QoP_S}$ and $\mathrm{QoP_U}$ are concerned, a similar link between both components of QoP is not characteristic of the other clips. The link between entertainment and content understanding is therefore not direct and represents a possible avenue for future investigations.

IV. Conclusions

This paper augments the traditional view of QoS with that of QoP, which is comprised of a user's satisfaction with the entertainment value of multimedia presentations together with the benefit of such presentations in terms of content assimilation and understanding. Our research has shown that a significant loss of frames or color depth reduction does not proportionally reduce users' understanding of and satisfaction with the presentation. This has important implications for bandwidth allocation in multimedia applications. Users also have difficulty in absorbing audio, visual, and textual information concurrently, tending to focus on one of the visual media and audio at any one moment. Lastly, user satisfaction, although strongly related to content, depends on the purpose of the presentation, as users are likely to ignore QoS degradations if also viewing presentations for informational

All these results indicate that the notion of quality in distributed multimedia systems must encompass user perceptual considerations if multimedia presentations are to be truly effective. A framework for an integrated solution to the delivery of multimedia data based on the results of this study has been proposed [5], and it is our belief that our work paves the way for truly end-to-end communications architectures, incorporating user perceptual requirements with networking considerations.

References

[1] R. T. Apteker, J. A. Fisher, V. S. Kisimov, and H. Neishlos, "Video acceptability and frame rate," *IEEE Multimedia*, vol. 2, no. 3, pp. 32–40, Fall 1995.

[2] G. Blakowski and R. Steinmetz, "A media synchronization survey: Reference model, specification, and case studies," *IEEE J. Select. Areas Commun.*, vol. 14, no. 1, pp. 5–35, Jan. 1996.

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

[4] K. Fukuda, N. Wakamiya, M. Murata, and H. Miyahara, "QoS mapping between user's preference and bandwidth control for video transport," in *Proc. 5th Int. Workshop on QoS*, New York, 1997.

[5] G. Ghinea, G. D. Magoulas, and J. P. Thomas, "Intelligent management of QoS requirements for perceptual benefit," in *Proc. 3rd Conf. Intelligent Systems Design and Applications*, Tulsa, OK, 2003, pp. 437–446.

[6] J. Kawalek, "A user perspective for QoS management," in *Proc. QoS Workshop Aligned with the 3rd Int. Conf. Intelligence in Broadband Services and Network (IS&N 95)*, Crete, Greece, Sep. 1995.

[7] R. E. Mayer, "Multimedia learning: Are we asking the right questions?," *Educat. Psychol.*, vol. 32, no. 1, pp. 1–19, 1997.

[8] T. Yamazaki, "Subjective video assessment for adaptive quality of service control," in *Proc. IEEE Int. Conf. Multimedia and Expo.*, Tokyo, Japan, Aug. 2001, pp. 517–520.

[9] 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, 1999.