ACM Multimedia 98 - Electronic Proceedings

QoS Impact on User Perception and Understanding of Multimedia Video Clips

G. Ghinea
Department of Computer Science,
University of Reading
RG6 6AY, Berks, U.K.
Tel. No. +44-118-9875123 x 7643
G.Ghinea@reading.ac.uk
J.P. Thomas
Department of Computer Science,
University of Reading
RG6 6AY, Berks, U.K
Tel. No. +44-118-9875123 x 7631
J.P.Thomas@reading.ac.uk

Abstract

The widespread and increasing advent of multimedia technologies means that there must be a departure from the viewpoint that users expect a Quality of Service (QoS) which will only satisfy them perceptually. What should be expected of multimedia clips is that the QoS with which they are shown is such that it will enable the users to assimilate and understand the informational content of such clips. In this paper we present experimental results linking users' understanding and perception of multimedia clips with the presentation QoS. Results show that the quality of video clips can be severely degraded without the user having to perceive any significant loss of informational content.

Keywords

QoS, multimedia, user perception, user understanding

Table of Contents

Introduction

Multimedia QoS is typically measured using technical parameters such as end-to-end delay and jitter. Such technical parameters, although useful, disregard the users' perspective of the presentation. The human element is an important part of the multimedia paradigm whose role, although appreciated, has often been overlooked. This is because of the inherent difficulty and subjectivity associated with appreciating an individual's sense of multimedia perception and, consequently, precious little work has been done in this area. What work [1], [2], [4] has been done, however, has indicated the existence of a threshold beyond which a user does not perceive an improvement in the QoS of multimedia applications, no matter the amount of resources allocated to them. Some research has also been done in order to establish the synchronisation limits between audio and video streams of a multimedia clip with which human observers are comfortable [3].

We are currently involved in investigating the impact that multimedia QoS has not only on a user's satisfaction with the quality of the presentation itself, but also on his/her capacity to understand, analyse and synthesise the informational content of such presentations.

The motivation behind this approach is that in multimedia it is not only the aesthetics that count, but also the effect of resource allocation on the user's potential to comprehend and assimilate the material and data presented in multimedia applications. This is especially important as multimedia databases become widespread and the technology is used in information-intensive domains such as education. Moreover, the QoS impact on user perception and understanding also has implications on network protocol design and resource allocation.

<-- Table of Contents

Outline of Approach

In our approach users were presented with a series of windowed (352*288 pixels) MPEG-1 video clips. If users habitually wore glasses and did not have their reading glasses on, they were told to put them on, as there might be instances where they might be asked questions about textual information displayed on the screen. Each of the clips, 12 in all, was between 31 and 45 seconds long. After the users has seen each clip once, the window was closed, and they were asked a number of questions about the video clip they had just seen. The actual number of such questions depended on the video clip, and varied between 10 and 12. After the user had answered the set of questions pertaining to a particular video clip and the responses had been duly noted, (s)he was asked to rate the quality of the clip that had just been seen on a scale of 1 - 6 (with scores of 1 and 6 representing the worst and, respectively, best perceived qualities possible). The user then went on and visualised the next clip.

Users were instructed not to let personal bias towards the subject matter in the clip or production-related preferences (for instance the way in which movie cuts had been made) influence their quality rating of a clip. Instead, they were asked to judge a clip's quality by the degree to which they, the users, felt that they would be satisfied with a general purpose multimedia service of such quality. Users were told that factors which should influence their quality rating of a clip included clarity and acceptability of audio signals, lip synchronisation during speech, and the general relationship between visual and auditory message components.

In our experiments, we have varied the frames per second (fps) QoS parameter. Parameters that were kept constant include the colour depth, window size and audio stream. Three frame rates were played: 25, 15 and 5 fps. Although these parameters varied across the experiments, for a particular user they were kept constant. The user furthermore was kept unaware of the frame rates at which the clips were being shown to him/her. 10 users were tested for each frame rate.

There are two main reasons why it was decided that the audio stream would be played with its original recorded parameters:

1. Bandwidth is the main resource we are interested in using more efficiently. As the audio stream occupies a very small bandwidth of the multimedia clip, compression at this level will not result in major gains in bandwidth.
2. It has been already shown that audio information has primacy over video content [3]. We were especially interested in the interplay between frame loss and assimilation of informational content

For each clip, the questions were chosen to encompass all aspects of the information - audio, visual or textual - presented in the clips. In addition to this, some questions could only have been answered if the user had grasped pieces of both visual and audio pieces of information from the clip. Other questions, as will be shown later, were also chosen to see what was perceived as being the feel or the atmosphere of the clip. In order to be confident that the results were based purely on variations in the frame rate, questions were asked immediately after each clip so that the information contained was still fresh in the memory of the participants. 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.

The clips themselves were chosen to cover a broad spectrum of subject matter in which the following factors were specifically taken into account:

- **spatial parameters**
- **temporal parameters**
- **importance of audio information in the context of the clip**
- **importance of the video information in the context of the clip**
- **importance of textual information in the context of the clip**

VIDEO CATEGORY	DYNAMIC	AUDIO	VIDEO	TEXT
1. Action Movie	2	1	2	0
2. Animated Movie	1	1	2	0
3. Band	1	2	1	0
4. Choir	0	2	1	0
5. Commercial	1	2	2	1
6. Cooking	0	2	2	0
7. Documentary	1	2	2	0
8. News	0	2	2	1
9. Pop Music	1	2	2	2
10. Rugby	2	1	2	1
11. Snooker	0	1	1	2
12. Weather	0	2	2	2

Table 1: Video Categories

Table 1 contains a brief characterisation of the clips used in our experiments. This was obtained by assigning weights on a scale of 0-2 corresponding to the importance of the audio, video and textual information in the context of the clip, as well as the dynamism of the clip itself.

As far as the subject matter itself is concerned, the clips were chosen such that they would present the majority of users with at least a minimal amount of interest, but not a great deal, in which case the answers given might possibly be skewed. As an example, when it was decided on what type of action sport video clip to present, we finally chose rugby, a sport which, although most Europeans are familiar with, does not enjoy the widespread following and close interest of football. It would have been more likely that a user would have benefited from his/her previous knowledge and experience of the game of football to answer correctly any questions that might have been posed.

Lastly, a few remarks shall be made about user behaviour during these experiments. Users were told that, although attention was required of them, they should view the clips in a relaxed manner, as the experiment was a data gathering exercise and not an exam. Needless to say, some users (probably among the more perfectionist ones) concentrated and thought for quite long periods of time when an immediate answer would not come to mind. What also happens in such experiments is that, after the first few set of questions have been answered, users will try and second guess the questions that might be put in subsequent clips that they see. The variety of the subject matter being shown (the fact that any subsequent clips seen will belong to different categories), as well as the fact that we had second guessing in mind when we formulated each set of questions ensured that this approach to answering questions was successful only for very few people and then only with a few isolated questions.

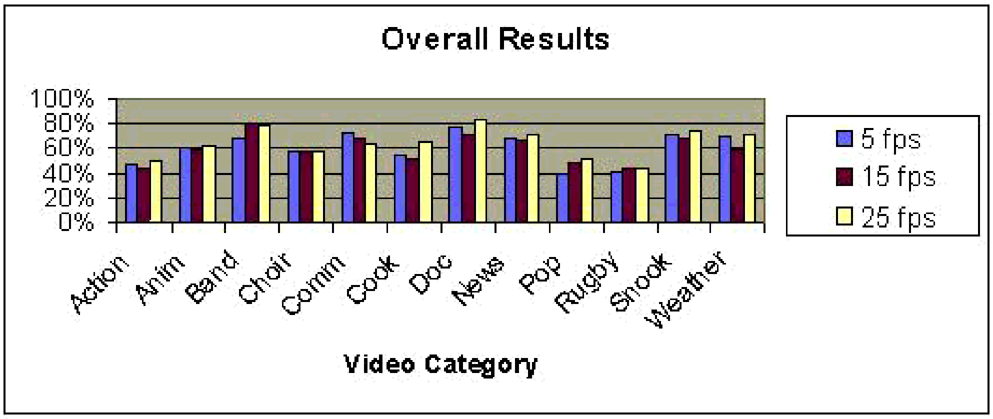
<-- Table of Contents

Experimental Results

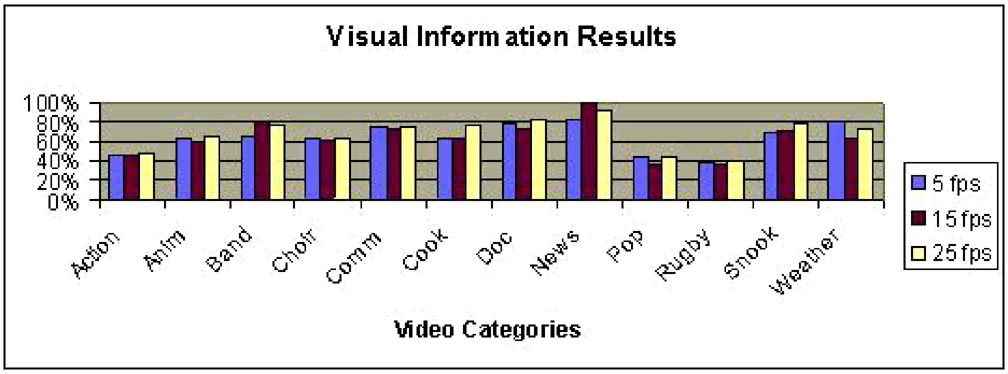
A few interesting remarks must be made about how the users answered some of the questions. The commercial video clip, for instance, is of a washing liquid for bathrooms and depicts a couple extolling its qualities. One of the questions which was asked was what the user thought the relationship between the couple was. The clue here was that there was a shot of the man's hand cleaning the bathroom in which a wedding ring could clearly be distinguished. What was, however, interesting was that 96% of the tested users said that the couple in the ad must be married. In all such cases, this answer was given not because the participants had seen the wedding ring, but because this was the way in which they had perceived the situation to be or they had felt that this was the target market being addressed.

Another remark needs to be made about the snooker video clip that was also part of our experiments. In this clip, the player pockets a red ball; however, because of the production lighting, the colour of the ball comes out as dark brownish. When asked what colour the pocketed ball was, very few people unfamiliar with the game of snooker actually said red. Those who enjoyed the game, however, always got the answer right, the motivation being that the player's current score (displayed as textual information on the bottom of the screen) had increased by the number of points corresponding to a red ball being pocketed.

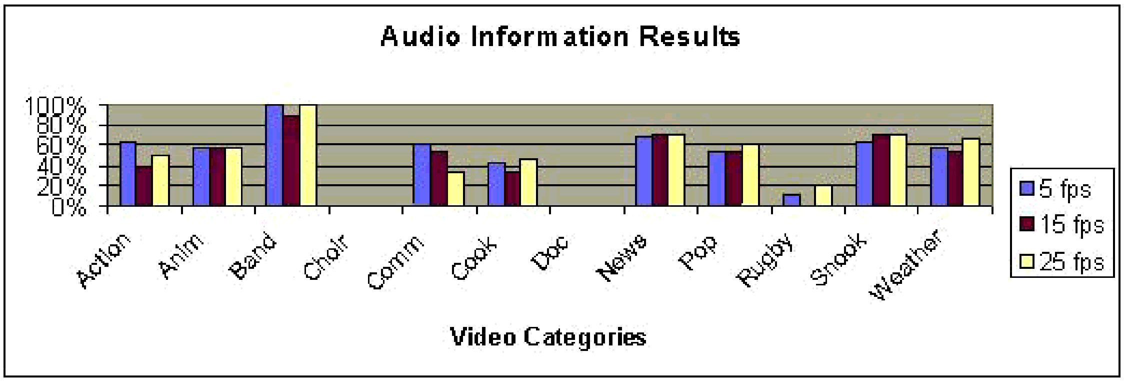
The last observation concerns the manner in which users answered questions regarding the cooking clip. Here, users were asked whether or not they had seen any forks in the clip. Most answered yes, which was the correct answer. However, the only forks that appeared in that clip were two very large ones hung on the walls for decorative purposes. This was observed by roughly a third (30%) of the respondents who answered correctly; the remaining correct answers were due to the fact that the users had assumed that, it being a cooking clip, there must naturally be forks somewhere along the line.



Our results show that there is no significant difference between the percentage of correct answers given by respondents at different video frame rates. This would seem to indicate that severe frame dropping does not have a proportional impact on users' capacity to assimilate video clip material. Indeed, in some cases, the percentage of correct responses was marginally higher at lower frame rates. This could be explained by the fact that, the complementary process of 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 (3 or even 5 times longer in the case of our experiments) on the screen. This would therefore increase the chance of the user noticing the respective information.



As expected, the lowest percentages of correct answers were given in action clips with rapidly varying scenes - the action movie, the rugby clip - or those boasting a rich diversity of informational content such as the pop clip. In the latter clip, for instance, in addition to the audio (of primary importance in this case) and video streams (where the body language and demeanour of the singer also tries to convey a message), textual information about the singer was regularly displayed on the screen. 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. In the case of informationally rich clips, what usually happens is that users cannot distribute their attention. For example, a frequent remark in the case of the pop clip was that "I was enjoying the music and wasn’t interested in the text", which would explain why respondents got such low percentages of correct answers in this case.



Since the audio stream is unaffected by frame dropping, one would initially expect that respondents would score much better when asked about the audio content of the clip. This is, generally speaking, the case. However, there are some exceptions: in the case of the rugby clip, when asked what particular feature (an ’overlap’, clearly mentioned by the commentator) of the attacking team had made the score of a try possible, only a handful of people got the answer; the majority of them were concentrating on the action of the game itself. Similar comments apply to the pop clip - when asked questions pertaining to the lyrics of the song, many people said that they hadn’t paid attention to the lyrics themselves, they were enjoying the melody in general. This happened especially when the clip was run at the full 25 fps - probably people were enjoying the overall quality of the clip, without giving regard to specifics.

As far as the satisfaction associated with media clips is concerned, a few observations need to be made. Generally speaking, the lower the frame rate is, the lower the user's satisfaction with it, although the variation is not linear. Users seem to have enjoyed the animated clip at the expense of assimilating the data; a similar remark can be made about the rugby clip. As far as users' perception of dynamism goes, this latter clip and the action movie received similar across the board satisfaction ratings. As concerns the news clip, users were annoyed at the newscaster's visible lack of lip synchronisation and thus, even though it was a static clip, only gave it average values as far as satisfaction is concerned. Users then essentially treated the bulletin as an audio broadcast, proof being the consistently high percentages of correctly answered questions related to the audio clip of this particular clip. Lastly, one can remark that in the case of information rich media such as the pop music clip, users tend to discriminate drastically according to the perceived quality of the presentation, there being a strong dependency between the perceptual satisfaction and the displayed frame rate.

<-- Table of Contents

Conclusions

Instead of the traditional technical definition of QoS, this paper defines QoS from the user's standpoint. In our view this is comprised of a user's perception of multimedia presentations together with the benefit of such presentations from a user's angle in terms of content assimilation and understanding. The main conclusions drawn from this work may be summarised as follows:

- A significant loss of frames (that is, reducing the frame rate) does not proportionally reduce the user's understanding and perception of the presentation. In fact, in some instances (s)he seemed to assimilate more information, thereby resulting in more correct answers to questions. This is because the user has more time to view a frame before the frame changes (at 25 fps, a frame is visible for only 0.04 sec, whereas at 5 fps a frame is visible for 0.2 sec), hence absorbing more information. This observation has implications on resource allocation.
- Users have difficulty in absorbing audio, visual and textual information concurrently. Users tend to focus on one of these media at any one moment, although they may switch between the different media. This implies that critical and important messages in a multimedia presentation should be delivered in only one type of medium, or, if delivered concurrently, should be done so with maximal possible quality.
- The link between perception and understanding is a complex one; when the cause of the annoyance is visible (such as lip synchronisation), users will disregard it and focus on the audio message if that is considered to be contextually important.
- Highly dynamic scenes, although expensive in resources, have a negative impact on user understanding and information assimilation. Questions in this category obtained the least number of correct answers. However the entertainment value of such presentations seem to be consistent, irrespective of the frame rate at which they are shown. The link between entertainment and content understanding is therefore not direct and this is further confirmed by the second observation above.

All these results indicate that Quality of Service, typically specified in technical terms such as end-to-end delay, must also be specified in terms of perception, understanding and absorption of content if multimedia presentations are to be truly effective.

<-- Table of Contents

Acknowledgements

Gheorghita Ghinea is sponsored in his work by the Reading University Scholarship Trust Fund. A thanks also goes to Roberto Fraile who helped in the set-up and execution of some of the experiments.

<-- Table of Contents

References

1. Apteker, R.T., Fisher, J.A., Kisimov, V.S., and Neishlos, H. Video Acceptability and Frame Rate. IEEE Multimedia, 2(3), Fall 1995, 32-40
2. Fukuda, K., Wakamiya, N., Murata, M., and Miyahara, H. QoS Mapping between User's Preference and Bandwidth Control for Video Transport, in Proceedings of the 5th International Workshop on QoS (IWQoS), New York, USA , May 21-23, 1997, 291 - 301
3. Kawalek, J. A User Perspective for QoS Management, in Proceedings of the QoS Workshop aligned with the 3rd International Conference on Intelligence in Broadband Services and Network (IS&N 95), Crete, Greece, 16 September 1995
4. Steinmetz, R. Human Perception of Jitter and Media Synchronisation. IEEE Journal on Selected Areas in Communications, 14(1), January 1996, 61 -72
5. van den Branden Lambrecht, C.J., and Verscheure, O. Perceptual Quality Measure using a Spatio-Temporal Model of the Human Visual System, in Proceedings of the SPIE, vol. 2668, San Jose, CA, January 28 - February 2, 1996, 450 - 461

<-- Table of Contents