High-bandwidth video conferencing systems: When is the quality worth the cost?

Emilee J. Rader
School of Information
University of Michigan
Ann Arbor, MI 48109 USA
ejrader@umich.edu

Erik Hofer
School of Information
University of Michigan
Ann Arbor, MI 48109 USA
ehofer@umich.edu

Tom Finholt
School of Information
University of Michigan
Ann Arbor, MI 48109 USA
finholt@umich.edu

ABSTRACT

The marketing literature for videoconferencing products often makes recommendations for settings that will yield acceptable subjective video quality; however, these recommendations have not been empirically validated. Additionally, a number of high bandwidth conferencing systems have been created that promise high video quality, but with extremely high bandwidth costs. An experiment was conducted investigating the subjective perception of video quality under the following conditions: 384 kbps, 1920 kbps, DVTS, and plain video. Four test scenes were used that differed in the amount and type of motion present in the video. Participants viewed sequences of reference and test scenes on a Samsung 42-inch plasma display, and made subjective ratings of all scenes using a discrete, 5-point scale. 1920 kbps scenes were given higher ratings than 384 kbps, but lower than both DVTS and a control. Ratings were also lower for scenes with more motion present in the video. No statistically significant difference was found between the DVTS and control test scenes. These findings suggest that while low bandwidth video may be acceptable for "talking heads"-type scene composition, medium bandwidth video offers a higher quality experience for users. Even higher bandwidth systems improve quality more, with the most dramatic improvements coming in scenes with high levels of motion.

Categories and Subject Descriptors

H.4.3 [**Communications Applications**]: *Computer conferencing, teleconferencing, and videoconferencing.* C.4 [**Performance of Systems**]: *Measurement techniques.*

General Terms

Measurement, Experimentation, Human Factors.

Keywords

Subjective assessment, video quality, videoconferencing.

1. INTRODUCTION

Video conferencing systems have been a major focus of telecommunications research, resulting in a large number of products and systems that can be used to transmit and receive video in real time over IP networks. These systems vary dramatically in terms of quality and in the amount of bandwidth used to transmit video over networks. The commercial sector has focused on developing low- to medium-bandwidth conferencing systems based on the H.320 and H.323 standards. These systems are able to operate using as little as 128 kbps of bandwidth, though 384 kbps is considered the standard connection speed for an acceptable conference [1]. Current generation H.323 endpoints are able to operate at higher bandwidths, many up to 2 Mbps, producing video with a higher frame rate and quantization. In addition to these commercial products, the academic research and higher education community has also pursued much higher quality and higher bandwidth systems that operate over high speed research and education networks, such as Internet2's Abilene Network. The Digital Video Transport System (DVTS) [2], for instance, IP-encapsulates DV25 video from an IEEE 1394 interface for transmission over IP networks, resulting in a high quality DV25 video stream that consumes roughly 30 Mbps of bandwidth. The bandwidth costs for these different systems vary dramatically, yet we lack a good understanding of how much difference higher bandwidth systems make relative to low or medium bandwidth systems in improving the subjective quality of video.

Because of the lack of understanding about the impact of these different technologies on subjective quality, many videoconferencing users accept the conventional wisdom that 384 kbps is all that is needed for a "successful" videoconference. Others, like the New World Symphony rely on the high bandwidth DVTS for music instruction over video in order to maximize quality [2]. The aggregate costs of systems like DVTS are significantly higher than low-bandwidth H.323 systems when bandwidth costs are factored in. Given the different quality choices and bandwidth costs of these systems, it is important to understand when and if the higher quality codecs will justify their bandwidth cost. Improving this understanding will better guide users in technology selection, in improving the quality of conferencing experiences, and in saving substantially on infrastructure costs.

Many factors have the potential to influence video quality. Gili et al. [3] identified seven factors relating to encoding and playback algorithms that influence quality. Due to the similarities in commonly used algorithms, three factors are generally understood to impact video quality among the systems heavily used for conferencing: frame resolution; frame rate; and frame quantization. Frame resolution refers to the number of pixels that make up a given frame in a video clip. Higher resolution frames are able to display more detail, resulting in a crisper image. Frame rate refers to the number of frames presented per second. NTSC television is composed of 30 frames per second, resulting in smooth video. Many studies and standards emphasize the frame rate as a key determinant of video quality. Frame quantization refers generally to the quantization factor of the discrete cosine transformation algorithm used in block-based video compression algorithms like JPEG, DV and the intraframe compression portions of H.261 and H.263. McCarthy et al. [4] recently identified quantization as the key determinant of quality in certain cases – more important than frame rate.

Accurate measurement of video or image quality is a challenge, though many products and metrics have been developed for quality assessment of television video. Internet-based video however, still relies on subjective experiments for quality evaluations because of the wide range of impairments that can result from different encoding techniques and formats. A typical subjective experiment involves having several non-expert viewers watch short clips of video shot from a constant camera angle and rate the quality of those clips on a 5-point Likert scale [5].

This paper discusses differences in perceived video quality of four test scenes under the following conditions: 384 kbps, 1920 kbps, DVTS, and NTSC playback of a DVD (control). Test scenes differed in the amount of motion they depicted. Data consisted of participants' subjective ratings of video quality immediately after viewing each test scene. Poor ratings were expected for high-motion scenes encoded at 384 kbps, and favorable ratings were expected for the control, and for scenes with less motion.

2. EXPERIMENT SETUP

Four test scenes were obtained from the Video Quality Experts Group [6]. Each scene was eight seconds in duration. We selected scenes that differed in the amount of motion depicted in the video, and in subject matter. Table 1 summarizes the visual content of the test scenes.

Table 1. Content of test scenes

Name	Description
Balloon	Man at amusement park holding a child while popping balloons. Motion of rides in the background.
Football	American football game where several players are running around on the field.
Phone	Close-up head shot of a woman talking on the telephone. Little motion.
Ship	Pirate ship floating in still water next to a dock. Almost no motion.

The clips were encoded from their original YUV format into MPEG2 for DVD playback. The MPEG2 clips were then output from a computer over an analog Y/C video interface to a set of hypothetical reference circuits and re-captured over an analog Y/C video interface and re-encoded as MPEG2. They hypothetical reference circuits included two H.323 endpoints connected at 384 kbps, resulting in a 320 kbps H.263 stream, two H.323 endpoints connected at 1920 Kbps, resulting in a 1872 kbps H.263 stream, and two DVTS endpoints connected at 30 Mbps, resulting in a DV25 stream.

3. METHOD

3.1 Experiment Design

A 4 x 4 x 4 mixed design experiment was conducted with 20 participants. Within-subjects independent variables were *encoding method*: 384 kbps, 1920 kbps, DVTS, and plain video (control), and *test scene*: balloon, football, phone, ship (see Table 1). Each participant rated all test scenes multiple times, encoded using each method. *Presentation order* of encoding methods was the only between-subjects independent variable. The order of presentation was completely counterbalanced, yielding four sequences of test scenes that were each presented to 5 participants.

3.2 Participants

Five men and fifteen women who were students and staff at a large university participated in the experiment. Two participants reported having previous experience with rating video quality. Six participants indicated that they did not wear glasses or contact lenses to correct their vision. The mean reported visual acuity of the 20 participants was 3.05 on a scale of 1 to 5, where 1 = below average, 3 = average, and 5 = above average.

3.3 Equipment

The display device used in this experiment was a Samsung 42" EDTV plasma display. Native resolution of the display was 852 x 480, and pixel pitch was 1.08x1.08(mm). Up to three participants at a time per experiment session were seated in front of the display (see Figure 1) such that the center chair was 63" from the center of the display, and distance to the left and right chairs was 67". Participants were not permitted to move their chairs during the experiment.

A 1.5 GHz Mac PowerBook G4 with 512 MB DDR SDRAM and 32 MB VRAM was connected to the plasma display's s-video port, and used to play the test sequences. The software used for playback of the video was Quicktime Player version 7. Test scenes were displayed at their original size.

Indirect incandescent track lighting was used to illuminate the room during the experiment, in an effort to reduce glare on the screen that might interfere with participants' ability to see the video clearly.

3.4 Procedure

The subjective evaluation procedure used in this experiment was based on ITU-R Rec. BT.500-11 Double Stimulus Continuous Quality Scale [5]. Participants were first shown each of the four test scenes in plain video format. These scenes were not rated, but were used to set a baseline for subsequent judgments.

Next, participants were told that they would be viewing four sequences of test scenes, and that some of the scenes had been modified and some had not. They were not told which were plain video "reference" scenes, and which had been encoded. In each sequence, scenes were presented in the following order: Balloon, Football, Phone, and Ship. Within each sequence, each scene was played twice, first as plain video (the "reference" scene), and a second time encoded at 384 kbps, 1920 kbps, DVTS or plain video was repeated. At the end of the experiment, participants had seen each clip eight times. There was a 3 second pause between each scene during which participants rated the video quality of the scene on a discrete five-point scale, where 1 = excellent and 5 = bad. There was a 6 second pause between sequences. Finally, a brief questionnaire was administered at the end of the experiment.

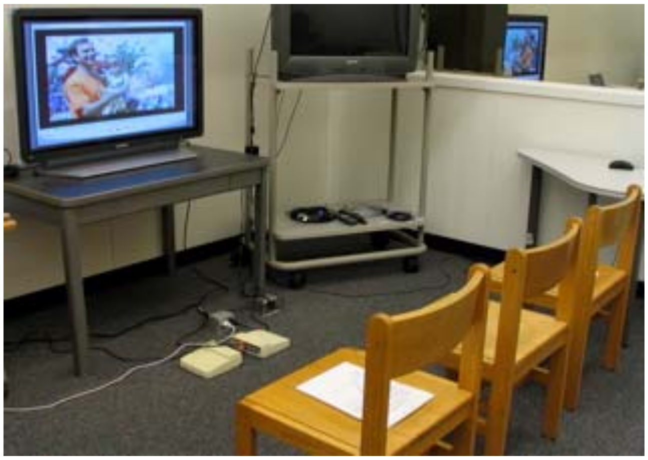

Figure 1. Experiment Setup

4. RESULTS

Data were prepared for analysis following the “hidden reference removal” procedure [7]. Participants were not informed which scenes in the sequences they viewed were encoded and which were unaltered “reference” scenes. Ratings for altered scenes were subtracted from the ratings for the corresponding reference scenes to yield difference scores. These difference scores were used in all subsequent analyses.

A three-way mixed design ANOVA (Encoding method x Test scene x Presentation order) revealed main effects of encoding method, F(3,48) = 85.597, p < .05, and test scene, F(3,48) = 26.224, p < .05. There was no significant main effect of Presentation order, indicating that the order in which participants viewed the encoding methods did not influence their ratings of the test scenes. The ANOVA also revealed a significant Encoding method x Test scene interaction, F(9, 144) = 11.958, p < .05. The sphericity assumption was not violated in any of the repeated measures analyses.

Table 2. Mean difference scores, Encoding x Test Scene

	384 kbps		**1920 kbps**		**DVTS**		**Control**	
	mean	sd	mean	sd	mean	sd	mean	sd
Balloon	2.05	.759	1.10	.788	.15	.587	.20	.616
Football	2.55	.826	.95	.605	.05	.605	.10	.553
Phone	1.40	.821	.90	.641	.40	.503	.15	.671
Ship	.20	.696	.20	.410	-.25	.639	.30	.571

Post hoc repeated measures ANOVAs were done to determine specific differences between test scenes and encoding methods. The alpha level was adjusted to .005 (.05 / number of comparisons) using the Bonferroni procedure to reduce the likelihood of Type I error due to multiple analyses.

As expected, the 384 kbps test scenes were given ratings that were lower than the other encoding methods (see Figure 2). However, the differences between 1920 kbps and DVTS were less clear-cut. A two-way repeated measures ANOVA (1920 kbps x dvts) revealed a main effect of Encoding method, F(1,19) = 53.971, p < .005. In addition, the comparison between DVTS and the control was not significant.

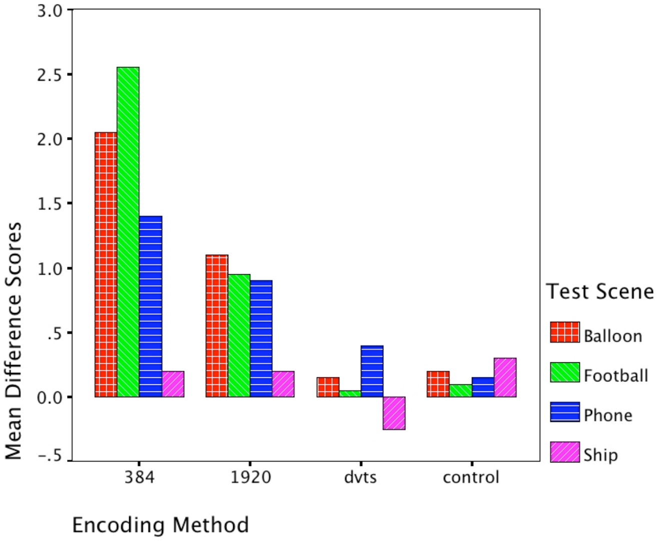

Figure 2. Encoding Method x Test Scene

Further two-way repeated measures ANOVAS were conducted to tease apart interaction effects. These results are summarized in Table 3, and show that the quality of the Phone scene is perceived more similarly to the high-motion Balloon and Football scenes than the Ship scene which contains almost no motion.

Table 3. Mean difference scores, Encoding x Test Scene

Comparison	**df**	**F**
Phone 384 kbps x Balloon 1920 kbps	1,19	.85
Phone 384 kbps x Football 1920 kbps	1,19	7.03
Phone 384 kbps x Ship 1920 kbps	1,19	31.81*
Balloon 1920 x Phone DVTS	1,19	15.26*
Football 1920 x Phone DVTS	1,19	12.84*
Phone 1920 x Phone DVTS	1,19	7.31

* p < .005

In addition, the Balloon and Football scenes at the 1920 kbps level were significantly different from the Phone scene at DVTS, but not the phone scene at 1920 kbps. This indicates that for scenes with even a small degree of motion, 1920 kbps is at a significant disadvantage.

In summary, DVTS test scenes with a lot of motion were perceived to be of equal quality to the control scenes. However, high-motion scenes were rated significantly worse than the control in the 382 kbps and 1920 kbps conditions. Clips with little motion were not rated significantly different from the control clips.

5. DISCUSSION

These findings provide some insight into the bandwidth/quality tradeoffs in video conferencing systems. The higher-bandwidth systems resulted in a higher perceived quality, but the differences were most dramatic in cases when there was a large amount of motion in the scene. This suggests that the higher-bandwidth cost is justified in cases where there is a lot of detailed motion in the scene. In a conferencing context, this commonly occurs when a scene involves a number of people instead of a close-up of one person.

That is, in cases where a conference involves a smaller number of people or only a single participant, the medium-bandwidth systems are likely acceptable, realizing a large quality increase over low-bandwidth systems – and a smaller quality decrease relative to high-bandwidth systems. This contradicts the widely held assumption that low bandwidth systems should be adequate for scenes limited to single face shots.

While this study does provide some insight into the quality and bandwidth tradeoff in conferencing system, it highlights a number of opportunities for future work that will strengthen the recommendations for users and developers of conferencing systems. One area for future research is to explore the impact of simulated network disturbances on the different codecs. Introducing network emulation into the hypothetical reference circuits will allow us to simulate real world conditions and make the findings more applicable to deployments of conferencing technologies. Another area for future development is introducing audio into the evaluation. Others [8,9] have looked at the interactions between audio and video quality in the overall quality of experience, suggesting that the audio component is also important in predicting quality. Finally, we would like to incorporate recent vision-based models [10] of quality into our experimental design to better anticipate at what point it makes sense to adopt a higher- or lower-bandwidth algorithm.

6. REFERENCES

[1] Trauner, M. and Yafchak, M.F. eds. (2005). *ViDe Video Conferencing Cookbook.* Lulu Press.

[2] DVTS Consortium. http://www.sfc.wide.ad.jp/DVTS/.

[3] Internet2 and New World Symphony. (2005). Internet2 and New World Symphony Performance and Master Class Production Workshop, Miami, FL.

[4] Gili, M.J., Janez, E.L., Hernandez, L.M. and Szymanski, M. (1991). Subjective image quality assessment and prediction in digital videocommunications. COST 212 HUFIS Report.

[5] McCarthy, J.D., Sasse, M.A., and Miras, D. (2004). Smooth or sharp? Comparing the effects quantization vs. frame rate for streaming video. In *Proceedings of the ACM Conference on Human Factors in Computing (CHI 2004)*, Vienna, Austria.

[6] ITU-R Recommendation BT.500-11, “Methodology for the subjective assessment of the quality of television pictures.” International Telecommunication Union, Geneva, Switzerland, 2002.

[7] VQEG. (2000). Final report form the Video Quality Experts Group on the validation of objective models of video quality assessment. http://www.vqeg.org

[8] Pinson, M. and Wolf, S. (2003). Comparing subjective video quality testing methodologies. In *Proceedings of SPIE ’03.*

[9] Watson, A. and Sasse, M.A. (1998). Measuring perceived quality of speech and video in multimedia conferencing applications. In *Proceedings of ACM Multimedia*, pp. 55-60.

[10] S. Winkler and C. Faller, "Audiovisual quality evaluation of low-bitrate video," presented at SPIE/IS&T Human Vision and Electronic Imaging, San Jose, CA, 2005.

[11] S. Winkler, *Digital video quality: vision models and metrics.* West Sussex, England: John Wiley & Sons, 2005.