Available online at www.sciencedirect.com
ScienceDirect
Decision Support Systems 45 (2008) 681–696
Decision Support Systems
www.elsevier.com/locate/dss

Internet-based telemedicine: An empirical investigation of objective and subjective video quality

Bengisu Tulu [a,*], Samir Chatterjee [b]

[a] Department of Management, Worcester Polytechnic Institute, USA
[b] Network Convergence Laboratory, School of Information Systems and Technology, Claremont Graduate University, USA

Available online 23 December 2007

Abstract

This study focuses on the Internet-based telemedicine with the goal of understanding the relationship between objective and subjective video quality measures and the decision making capability of medical professional using an ophthalmology video. Objective and subjective measures are calculated using PSNR and the perception of human viewers respectively. An emulated Internet testbed was created for experiments. Results indicate that jitter and delay have significant effect on the objective quality values. Subjective quality, on the other hand, not only depends on the same two factors, but also depends on which critical frames the provider is able to see and work with.

Keywords: Telemedicine; Quality measurements; Network impairments; Subjective quality experiments

1. Introduction

Telemedicine and e-Health in particular can spread critical medical expertise and health services across a region and around the globe. However, quality issues combined with cost and accountability are inhibiting the growth [3]. Even though telemedicine has tremendously evolved over the past 30 years, today most telemedicine implementations still require expensive leased telecommunication circuits to provide secured reliable connections. Studies are still having difficulties in finding evidence regarding the cost-effectiveness and practicality of telemedicine programs [20]. It is recommended [20] that regardless of the glamorous technology available, the primary aim of telemedicine should be to ensure that the most appropriate technology is used the most effective way.

Telemedicine applications rely on the telecommunication infrastructure which is often chosen carefully to support such applications. Internet and private IP-based networks provide a ubiquitous, standardized system interface [22] at low cost, and hence they can help deliver telemedicine services to a wider population. However, the unreliable connection properties of packet-based systems and their vulnerability to various impairments that affect the physical, network, and application layers hamper the quality of Internet-based telemedicine applications. Regardless of the transmission technology used in a telemedicine program, there is one single requirement for real-time telemedicine multimedia applications, that is, to provide the same quality before and after the transmission of packets over the telecommunication channel. This becomes incredibly hard when the Internet is used as the delivery channel.

* Corresponding author. Department of Management, Worcester Polytechnic Institute 100 Institute Road Worcester, MA 01609-2280, USA. Tel.: +1 508 831 5184; fax: +1 508 831 5720.
E-mail address: bengisu@wpi.edu (B. Tulu).

0167-9236/$ - see front matter © 2007 Elsevier B.V. All rights reserved.
doi:10.1016/j.dss.2007.12.009

This paper has two specific goals. The first goal is to explore the effects of certain network impairments (packet delay, jitter, and packet drop) that occur over the Internet on telemedicine video quality and to identify cutoff points for Internet-based telemedicine videoconferencing applications where video quality stays above an acceptable objective measurement level. The second goal is to understand the relationship between objective and subjective video quality measures and clinical decision making capability of medical professionals on the receiving end, when the Internet is the transmission channel.

Existing video quality measures, both objective and subjective, were originally developed for broadcast networks. The Internet has very different characteristics compared to these old traditional communication technologies. For example, Internet is a best-effort lossy network. Wolf and Pinson [23] stated that, "To be accurate, digital video quality measurements must be based on the perceived quality of the actual video being received by the users of the digital video system rather than the measured quality of traditional video test signals (e.g., color bar)". Especially in telemedicine, perceived video quality plays a critical role in the medical professional's confidence level in decision making, and hence quality within the context of telemedicine requires special attention. The ultimate goal of this research is to address the following questions:

> *"How can a medical practitioner conducting telemedicine over the Internet assign a metric to the quality of video received?"*

To answer this question, the paper is outlined as follows. First, a review of network impairments and their impact on video quality is provided followed by a summary of existing objective and subjective video quality measurements and quality research conducted in telemedicine. Second, the paper follows with the explanation of the two distinct phases of the experimental research design. In phase 1 objective measurement experiments are described including the study testbed, the experimental design, the data collection, the results and the data analysis. In phase 2 subjective measurement experiments are described including the experiments conducted using human subjects, the measurement methods utilized, the data collection, and the results. Finally, the paper concludes with a discussion of results obtained from the experiments and future research.

2. Background

2.1. Network impairments

The Internet Protocol (IP) is a packet-based network protocol that enables the transmission of data packets, from one end system to another based on address information carried in the packet. It can be used with two different transport layer protocols: Transmission Control Protocol (TCP) and User Datagram Protocol (UDP). TCP is a connection oriented, reliable transport protocol designed for data transmission. However, it is not suitable for real-time applications because the retransmission of packets may cause high delay and increase delay variation, which can significantly affect the quality of real-time applications. Therefore, real-time applications use UDP, a connectionless transport layer protocol, even though it does not guarantee the arrival of a packet.

Real-time multimedia applications[1] also utilize two protocols that run over UDP: the real-time transport protocol (RTP) and the RTP control protocol (RTCP) [19]. RTP is designed to carry data that has real-time properties. RTCP is designed to monitor the quality of service and to convey information about the participants in an on-going session. Even though RTP is the commonly used protocol for real-time applications; RTP, by design, does not provide any mechanism to ensure timely delivery or provide other quality-of-service guarantees, but relies on lower-layer services to do so. Therefore, real-time multimedia applications are vulnerable to any impairment that can happen in the lower layers of the network. These impairments may be due to lack of guarantee in terms of bandwidth, packet loss, packet delay, and jitter. All of these can affect the quality of voice and video over the Internet as reported in various studies [10,14,18].

2.1.1. Packet loss

Unlike circuit-switched networks, in packet switched networks no physical end-to-end circuit is established [10]. Packets are transmitted from the source to the destination over the Internet with the help of routers. Arriving packets at a router are first queued and then transmitted one-by-one, usually with the first in first out (FIFO) policy. However, if the queue (buffer) of a router is already full when a packet arrives, then this packet is dropped and consequently, is not

[1] While advanced network QoS techniques are available, these are very expensive and often not an option in rural and poor telemedicine regions.

transmitted to its destination. Network congestion occurs when routers start dropping packets. The effects of packet loss on real-time multimedia applications are critical and the effect of extensive packet loss on video is acute. If packet loss happens, some parts of the video cannot be decoded and displayed. It is important to understand the effects of packet loss on the perceived quality of voice and video applications.

2.1.2. Packet delay

End-to-end packet delay is typically caused by a number of components [10]: (1) codec delay is the time it takes to convert analog data to digital and vice versa, (2) serialization delay is the time it takes to place a packet on the transmission line, and is determined by the speed of the line, (3) queuing delay occurs at the various switching and transmission points of the network, such as routers and gateways, where packets wait in the queue to be transmitted over the same outgoing link, and (4) propagation delay is the time required by signals to travel from one point to another, which is fixed as determined by the speed of light. The effects of large packet delay become even more severe for voice communications, as timing is an important characteristic of voice. This is especially true when an interactive conversation is being transmitted on the network; delay effects can turn the conversation into a half-duplex mode where one party speaks and the other party listens and pauses to make sure the speaking party is done. Echo is another unwanted effect of packet delay.

2.1.3. Packet delay variation (Jitter)

Packet delay variation refers to the variation or gaps between packet arrival times at the receiving buffer. This occurs due to the variability in queuing and propagation delays. To eliminate the effects of this variation, usually a playout buffer is used. The receiver holds the first packet in the buffer for a specific amount of time before playing it out. Therefore, a small jitter is tolerable but large fluctuation causes difficulty in decoding and playback and causes quality degradation. The effects of delay variation are theoretically similar to the effects of packet loss. Large variation in delay will result in some packets arriving long after the playout time scheduled for them based on the buffer size. The receiver will discard these packets since they are out of order.

2.2. Video quality measurement

Quality measurement can be done either objectively (using electrical measurements) or subjectively (using human viewers) [21]. Peak Signal-to-Noise Ratio (PSNR) is the most commonly used objective metric for measuring video and image quality. It measures how close a sequence is compared to the original one [15]. The calculation of the PSNR for a video sequence of K frames each having NxM pixels with m-bit depth is calculated as explained in Eq. (1) [15]. First, the Root Mean Square Error (RMSE) is calculated according to the following formula:

$$RMSE = \sqrt{\frac{1}{N.M.K}\sum_{k=1}^{K}\sum_{n=1}^{N}\sum_{m=1}^{M}[x(i,j,k)-\bar{x}(i,j,k)]^2}$$

Equation 1

RMSE Formula for PSNR Calculation

where $x(i, j, k)$ and $\bar{x}(i, j, k)$ are the pixel luminance value in the i, j location in the k frame for the original and distorted sequences respectively. Once the RMSE is calculated, the PSNR can be calculated using the following formula:

$$PSNR = 10.\log\frac{m^2}{RMSE^2}$$

Equation 2

PSNR Formula

The PSNR is reported in decibels (dB) [18]. According to broadcasting standards, an image with a PSNR of 25 dB or below is usually unacceptable. Between 25 dB and 30 dB, perceived quality improves. Above 30 dB images are often perceived as good as the original image. It was noted by [13] that the PSNR is exclusively used as a quality measure, partly because of its mathematical traceability and partly because of the lack of better alternatives. On the other hand, it has also been noted by [15] that the PSNR does not always correlate well with subjective measures.

One other commonly used objective metric is the Video Quality Metric (VQM) [23], which was developed by the Institute for Telecommunication Sciences (ITS). It is designed for bench-top laboratory testing and is available for the PC and selected UNIX platforms. The tool implements video calibration algorithms (i.e., spatial registration, valid region estimation, system gain and level offset, temporal registration), root cause analysis algorithms (i.e., calibration problem detection, video artifact detection), and five video quality models (i.e., TV, Videoconferencing, General, Developer, PSNR). Details about the algorithms used in VQM can be found at Ref. [16]. It requires the extraction and classification of features from both the original and processed video sequences similar to the other measurement techniques. Once these features are extracted, the distance between the original and processed video sequences is computed based on these features; and later this distance is mapped to a subjective score [23]. Compared to the typical PSNR, this metric offers different models for various transmission types, and it is also possible to identify the nature of an impairment using the VQM [13].

The ITU-R 500 is the standard for subjective assessment of image quality and has evolved over the years to include measures for digital video transmissions as well. This standard provides scales for single and double stimulus methods. The Absolute Category Rating (ACR) is a single stimulus method where test sequences are presented one at a time and are rated on a category scale after they are viewed. Usually a 5-point category scale is used as illustrated in Table 1.

The Single Stimulus Continuous Quality Evaluation (SSCQE) is different from the ACR in terms of the scale it uses and the assessment process. Among the double stimulus methods, the Double Stimulus Impairment Scale (DSIS) — also known as the Degradation Category Rating (DCR) — presents pairs of original and impaired video sequences during the test respectively. In this case, subjects are asked to rate the impairment of the second stimulus with respect to the reference (first stimulus) using the 5-point impairment scale.

In the Double Stimulus Continuous Quality Scale (DSCQS) method, the sequences are presented in pairs like in the DSIS and subjects are asked to evaluate the quality of both sequences. The original sequence is included for reference; however, the observers are not told which one is the reference sequence and the order of appearance changes for each test. There are other methods where the two sequences are shown simultaneously and the observers are asked to make a comparison of the two based on stimulus comparison scale.

2.3. Video quality studies in telemedicine

Quality, in a telemedicine instance, can be measured at multiple points using various methods and measurements. Therefore, the literature of quality studies in telemedicine domain usually reflects different perspectives. As a common way of assessing quality of a telemedicine event, user satisfaction is used in a large number of articles. Another approach common in literature is to study the quality of the transmitted media (image, audio, video, etc.). These studies have been usually limited to the compression techniques and their effects on the perceived quality of the users. For example, Ref. [7] investigated image compression of digital retinal images and the effect of various levels of compression on the quality of the images. They compared JPEG and Wavelet image compression techniques and concluded that; "for situations where digital image transmission time and costs should be minimized, Wavelet image compression to 15 KB is recommended, although there is a slight cost of computational time. Where computational time should be minimized, and to remain compatible with other imaging systems, the use of JPEG compression to 29 KB is an excellent alternative".

To answer the question of which compression technique is better in a generic way, some studies focused on quality measures. An early study [4] considered an interesting question, "How does one decide if an image is good enough for a specific application, such as diagnosis, recall archival, or educational use?" and compared and contrasted three approaches to the measurement of medical image quality: the signal-to-noise ratio (SNR), a subjective rating, and diagnostic accuracy.

Table 1
ITU video quality assessment

5-point quality scale		5-point impairment scale	
Estimated quality	Score	Estimated impairment level	Score
Excellent	5	Imperceptible	5
Good	4	Perceptible	4
Fair	3	Slightly annoying	3
Poor	2	Annoying	2
Bad	1	Very annoying	1

They concluded that there is a need for computable measures of image quality that can accurately predict the outcomes of image quality evaluation studies. Another, more recent, article on image quality [17] stated that, "A numerical measure, which is able to predict diagnostic accuracy rather than subjective quality, is required for compressed medical image assessment." A new vector measure for image quality, reflecting diagnostic accuracy was developed in this study [17].

A recent study [18] focused on understanding the impact of variables affecting the transmission of video over IP networks. This recent study was one of the few studies that investigated the effects of network impairments and the codec bit rate on the quality of video on IP networks for telemedicine purposes. PSNR and a proprietary objective measurement technique, the Picture Quality Rating (PQR), were utilized. The reported findings suggest that an increase in codec bit rate and network bandwidth have positive effects on the PQR and the PSNR levels for sequences subjected to delay and jitter impairments, but not for those in which periodic packet drops were introduced. The results of this study indicated that with or without the existence of selected packet-specific impairments, increases in bandwidth and codec bit rate improve the objective quality of video transmitted over IP networks. Another study [6], which conducted measurement of perceived performance as a function of network delay, reported that the perception of the user regarding performance loss on an assigned visual task underwent degradation with increasing network delay only after the delay times exceed 100–200 ms.

Another study [5] presented a method to obtain an end-to-end characterization of the performance of an application over a network, by taking into account network impairments and application constraints. The applications selected for testing were two medical education tools: (1) an image serving application that delivers a sequence of linked images based on user movement of the mouse cursor and (2) an application intended to train students remotely in various surgical procedures. They were tested on four different types of networks. They stated that the subjective evaluations used in their study can be utilized to predict the conditions under which the application will be running based on predefined requirements.

3. Phase 1: objective measurement experiments

Understanding the effects of impairments, which occur in IP-based networks, on video quality in low-cost telemedicine settings requires the collection of data about the network conditions as well as the original and transmitted video sequences. Such data collection can be done in two ways: (1) sending packets over the real-world Internet and relying on the Internet service provider to collect the network data or (2) setting up an experimental testbed where Internet behavior is emulated to control network conditions and video sequence is transmitted and received over the emulated testbed. The former method provides more accurate data, and reflects the real distribution and combination of network impairments that occur on the Internet [8]. However, such experiments are both labor intensive and costly [8] since involvement of large service providers is necessary. The latter, on the other hand, is a feasible solution which combines the strengths of simulation and a real-world testbed.

> *"Emulation represents a comparatively recent effort to address the deficiencies of simulation through real-world interaction while retaining its strengths (repeatability and ease of configuration)* [8]*."*

Therefore, this study built an emulated testbed to conduct experiments in a controlled fashion. The following sub-sections explain the video sequence, the experimental testbed, the experiment design, and protocol.

3.1. Telemedicine video sequence

The telemedicine video sequence selected for the experiments was obtained from the Regenstrief Institute for Health Care at Indiana University. This video sequence and others were collected under the "Indianapolis Network for NGI Application to Telemedicine" project. Under this project, on-call, off-site physicians were able to conduct unscheduled videoconferencing with residents (patients) of a nursing home to assess acute medical problems occurring at night. The video sequence selected for this study was received through the National Library of Medicine and the patient shown in this video has provided informed consent for the video to be available to the general public.

The original video was in mpeg video format and 2 min 47 s long. Fig. 1 shows seven snapshots from the original sequence. It was a telediagnosis case in ophthalmology where an off-site, on-call physician conducted a general eye examination for the nursing home resident.

The mpeg video file (originally 800 pixels × 720 lines) was resized to 352 pixels × 288 lines, the Common Intermediate Format (CIF) size, based on the telehealth technology guidelines [9] where technology standards were first defined for teleophthalmology. These standards were categorized under different purposes of ophthalmology examinations. According to the purposes identified in these standards, the video used in our experiments presents an eye examination for external assessment. The technology guidelines for real-time external assessment

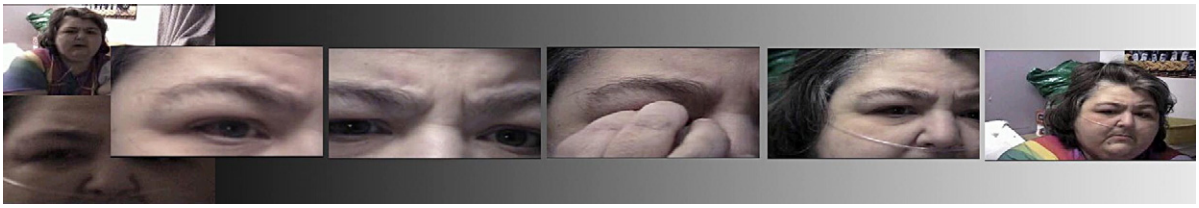


Fig. 1. Original video snapshots.

indicate that CIF image size is suitable for this type of examinations.

Original mpeg video was first downgraded to CIF size and converted into avi format since the VQM tool used requires video sequences in avi format. Later, this avi file was edited to create a video sequence of 14 s and black frames are included at the beginning and at the end of the file to eliminate any loss as a result of initialization or termination problems. This new file will be referred to as the *original sequence* from this point on.

3.2. Experimental testbed

The problem under investigation is the measurement of degradation in quality of video after it is transmitted through the Internet where impairments occur. A testbed, illustrated in Fig. 2 is designed to emulate the Internet traffic and control impairment parameters (delay, jitter, and drop) while transmitting an actual telemedicine multimedia session.

The testbed is composed of 5 components. There are two laptops (Client 1 — P4 CPU 1.8 GHz 256 MB, Client 2 — P4-M CPU 2.2 GHz 512 MB) with Microsoft Windows XP operating system, and running JM Studio for audio/video transfer using RTP, and Ethereal for network packet analysis. JMStudio [11] is a java-based media player developed based on Java Media Framework API. It can capture, play, record audio/video files. JMStudio can also receive and play RTP media streams. A Red Hat 9 Linux router running NIST Net is utilized to emulate Internet traffic and behavior. NIST Net is a network emulation package that allows a single Linux box to behave as a router to emulate a wide variety of network conditions, such as packet loss, duplication, delay and jitter, bandwidth limitations, and network congestion [1]. This allows testing of network-adaptive protocols and applications in a lab setting. The Linux router accommodates two network interfaces which are connected to the two local area networks where the two clients are hosted. Ethereal program runs on both network interfaces to monitor the network traffic and control the accuracy of the emulation happening at the router level.

3.3. Experiment design

In experimental design each variable that affects the response variable and has several alternatives is called a factor. Quality of video depends upon several factors. The experiments in this study are designed to isolate the effects of each factor from the effects of others so that meaningful results can be obtained. Proper experimental design allows determining if a factor has a significant effect or if the observed difference is simply due to random variations caused by measurement errors and parameters that are not controlled [12].

Video codecs and codec related parameters are controlled during the experiments. H.263 is an ITU video-coding standard originally designed for low bit rate communications (less than 64 kbits/s — this limitation has now been removed). It uses a similar coding algorithm than H.261 with some changes to improve the performance and error recovery. As a result of these improvements, H.263 output stream is more resilient to packet loss, which makes it very attractive for real-time communications over

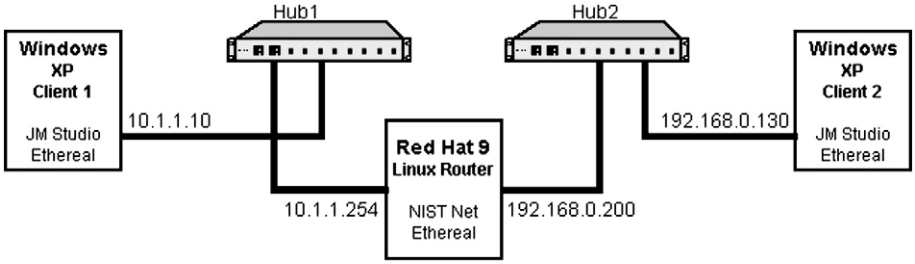


Fig. 2. Experimental testbed.

Table 2
Factors and their initial levels

Factors	Level 1	Level 2	Level 3	Level 4	Level 5
Delay (ms)	50	100	200	300	400
Delay Variation/ jitter (ms)	0	2	5	10	25
Drop (%)	0	–	–	–	–

the Internet. It supports five resolutions. In addition to CIF and Quarter CIF (QCIF), it provides resolution at Sub-Quarter CIF (SQCIF — 128×96 pixels), 4CIF (704×576 pixels), and 16CIF (1408×1152 pixels). Therefore, H.263 video codec with CIF (352×288) video size was selected for the experiments.

General full factorial design was selected with three factors. Table 2 presents the three factors (i.e. variables that affect the video quality) used in this experimental design and their initial levels (i.e. how they are manipulated). These levels are selected based on the previous studies of Internet backbone behavior [14]. In order to capture experimental errors, 2 repetitions for the first 25 experiments were conducted and the experimental error was calculated. This experimental design required 5 * 5 * 2 = 50 initial experiments including the repetition factor.

Results of the initial experiments were analyzed and based on the findings the factor levels were adjusted as presented in Table 3. After the adjustment, 3 * 5 * 4 = 60 experiments were required by the factorial design. The repetition factor was also excluded for the reasons explained in the results section.

In the experimental testbed illustrated in Fig. 1, the original sequence was stored on *Client 1*. JMStudio was started on *Client 1* and the *Transmit* option was used to send the original sequence using H.263 over RTP. On the receiving side, *Client 2* was running JMStudio with an RTP session on the port *Client 1* was transmitting video. Once the RTP session was established, the *Export* option of JMStudio was used to store the transmitted file on *Client 2* in avi file format with YUV video color option.

Meanwhile, NIST Net was set on the Linux Router to emulate a network based on different levels of factors in the experimental design (see Fig. 3). Details about the architecture and design of NIST Net can be found in Ref. [2]. In NIST Net, first Source and Destination fields were filled with the IP addresses provided on Fig. 2. Later, for each experiment, Delay, Delay sigma (i.e. delay variation), and Drop fields were used to control the network impairment levels between the source and the destination IP/Port combinations entered to the tool. Bandwidth was another controlled variable and it was set to 10 Mbps. Before transmitting the video, the emulator was turned on and the transmission of the video was started. After each experiment, the emulator was turned off to avoid accumulation of packets on the emulator and the queue was cleared. Traffic on all the network interfaces in the experimental testbed was analyzed using Ethereal during the transmission of the multimedia file.

The response variable used in this study is PSNR. Each experiment generated one processed video file (in total 110 degraded video files). Once objective measurements were completed, the amount of degradation based on PSNR was calculated using VQM tool.

4. Phase 1: results and data analysis

4.1. Initial experiments

The first step in the data analysis was calculating the PSNR values for all 25 processed video sequences generated during the initial experiments. In order to create a comparable video sequence, the original sequence was transmitted without any impairment on the testbed and stored as the reference processed video for PSNR calculations. During the initial experiments, delay and jitter were the only two factors that were manipulated (at 5 different levels each) in the emulator while drop percentage was set to 0. All 25 video sequences were compared against this reference video to analyze the level of degradation on video quality as a result of jitter and delay. These results are presented in Table 4. Columns represent different levels of jitter factor and rows represent different levels of delay factor. Each cell in this 5×5 matrix is filled with the PSNR values that were calculated by the VQM tool for the first repetition.

Tables 5 and 6 illustrate the column and row effects for Table 4 respectively. The results of the analysis indicate an average PSNR of 17.78 dB, which is in the unacceptable zone for video quality. Column effects, which represent the jitter effect, have higher impact on the PSNR value change compared to row effects, which represent the delay. Among different levels of jitter levels, no jitter has a −7.33 effect on the average PSNR value. No jitter also produces the only column mean (25.11 dB) that is in the acceptable range for quality.

ANOVA analysis of the initial experiment results is provided in Table 7. The table indicates that 80% of the

Table 3
Adjusted values of factors

Factors	Level 1	Level 2	Level 3	Level 4	Level 5
Delay (ms)	0	50	400	–	–
Delay Variation/ jitter (ms)	0	2	5	10	25
Drop (%)	0	5	10	25	–

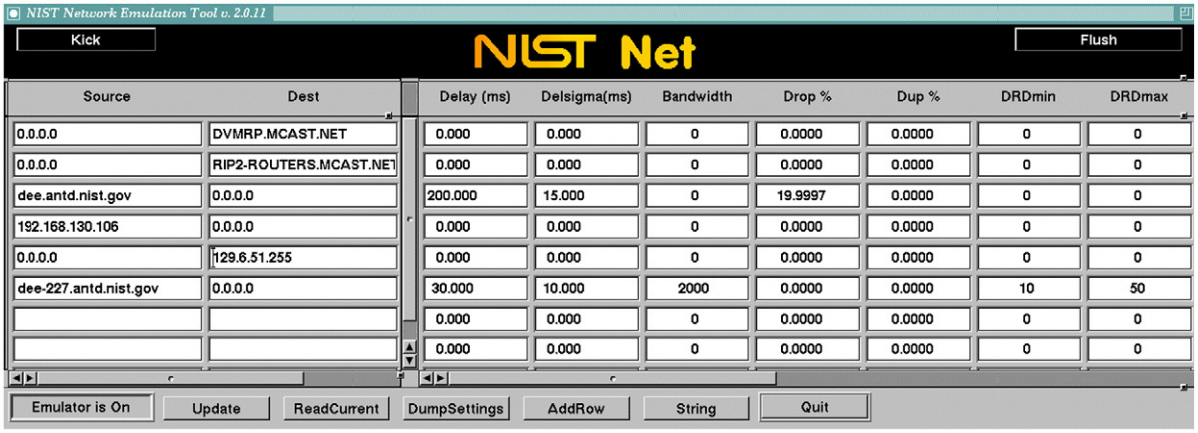


Fig. 3. NIST net network emulator user interface [23].

variance is explained by these variables. Out of two main effects, analysis is significant only for jitter. Both delay and interaction between delay and jitter did not provide any significant results. At this point, one can conclude that jitter is an important parameter in measuring video quality. F-ratio for jitter is higher than that obtained from the F-table. This confirms the previous conclusion that the size of jitter would make a significant difference in the video quality. On the other hand, the F-ratios for the delay and delay $*$ jitter are less than that obtained from the F-table. This shows that neither delay nor the interaction between delay and jitter have any significant impact on the video quality. Based on the conclusions of the initial experiments, the delay levels were dropped down to three (delay = [0 ms, 50 ms, 400 ms]) to lower the number of experiments that will be conducted in the next step.

Fig. 4 provides a graphical representation for the original PSNR values. This figure also supports our conclusion regarding the effects of jitter having a significant effect on the degradation of video quality.

4.2. Experiments with revised factor levels

Factor levels for the revised set of experiments were shown in Table 3 above. The repetition factor was dropped which led to one observation per condition. The results of the ANOVA, which is used to understand the main effects of the factors, are presented in Table 8. The results were similar to the findings of the preliminary test analysis. The effect of jitter was significant at 0.01 level and the effect of delay was significant at 0.05 level. This is different compared to the previous experiments where only delay and jitter was introduced as a factor. The third factor, drop, was expected to have similar effects as jitter. However, in our experiments, drop did not have any significant effect on the variance of PSNR values. Even though the R^2 was lower than the R^2 of the preliminary experiments; we observed an increase in the adjusted R^2 (.70) after including drop as a new factor in the model.

These results point out that the degradation in the video quality is mainly caused by the jitter (variation in delay), and packet delay. Drop did not have any effect on this degradation. Less significant effect of delay may be caused by the experimental setup. The video is streamed from one end to another and no interactive media component was utilized in the experiments. Lack of interactivity might have resulted in higher tolerance for delay because there was no need for synchronization between the two end points.

Based on the results of the experiments, the following conditions should provide a PSNR value of

Table 4
PSNR values for 25 experiments

(ms)	0	2	5	10	20
50	23.31	19.98	15.78	7.47	21.40
100	28.00	19.96	17.50	15.16	15.66
200	28.04	22.10	14.78	16.11	7.75
300	22.93	19.52	18.94	12.07	12.66
400	23.29	22.62	14.71	10.67	14.05

Table 5
Column effects

Jitter	0	2	5	10	20
Column sum	125.56	104.18	81.71	61.49	71.51
Column mean	25.11	20.84	16.34	12.30	14.30
Column effect	−7.33	−3.06	1.44	5.48	3.48

Table 6
Row effects

Delay	Row sum	Row mean	Row effect
50	87.94	17.59	0.19
100	96.27	19.25	−1.48
200	88.78	17.76	0.02
300	86.12	17.22	0.55
400	85.34	17.07	0.71

25 dB or more which corresponds to acceptable video quality:

(1) No delay variation/jitter, no packet drop, up to 400 ms delay;
(2) No delay, no delay variation/jitter, up to 5% packet drop.

5. Phase 2: subjective measurement tests

This phase builds on the previous one and measures the subjective quality of the degraded videos generated for objective measurements. Subjective quality experiments involve human subjects. When measuring video quality, the selection of subjects is usually based on their expertise in video quality measurement. However, in this case, the goal is not only to measure the video quality, but also to assess the clinical decision making capability based on the video under evaluation.

5.1. Sample selection

Subjects for this study were selected from different user groups. Since domain expertise is required for making a judgment on the conditions for a clinical decision after viewing the eye examination video, eight optometrists were recruited from public and private optometry clinics. Subjects with no domain knowledge were also included in the study to identify the different perceptions based on the background knowledge. Seven Information Systems and Technology (IST) doctoral students were recruited as the second group. The total sample size for this study was 15. ITU-T P.910 recommends having at least 15 participants who are not directly involved in picture quality evaluation as part of their work and are not experienced assessors. The sample satisfies ITU recommendations. However, the number of participants in this study is low compared to other studies in the field of Information Systems. This limitation may have effects on the statistical analysis.

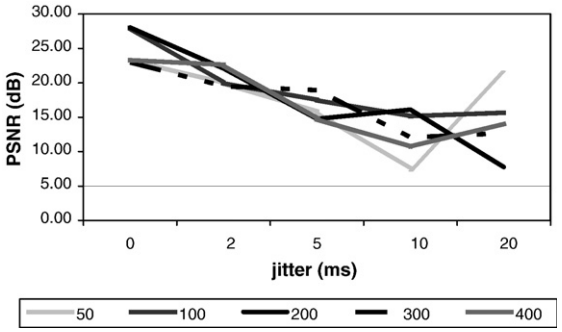


Fig. 4. Graphical representation of the results.

5.2. Selection of impaired video sequences

Fourteen impaired videos were selected from the pool of videos generated in the previous study. The goal was to generate a sample of video sequences where variation of all the three impairment factors and PSNR values were represented. Table 9 provides a list of 14 impaired video sequences and the original sequence with their PSNR values and impairment factors used to generate them.

5.3. Experimental design

Each subject was asked to fill out a short questionnaire about their expertise and their previous telemedicine

Table 7
ANOVA tests (dependent variable = PSNR)

Source	Sum of squares	*df*	Mean square	*F*	Sig.
Corrected model	2203.919[a]	24	91.830	4.351	.000
Intercept	17845.761	1	17845.761	845.635	.000
Delay	62.541	4	15.635	.741	.573
Jitter	1907.472	4	476.868	22.597	.000
Delay * jitter	233.907	16	14.619	.693	.775
Error	527.584	25	21.103		
Total	20577.265	50			
Corrected total	2731.504	49			

[a] *R* squared = .807 (Adjusted *R* squared = .621).

Table 8
ANOVA tests of between-subjects effects (dependent variable = PSNR)

Source	Sum of squares	*df*	Mean square	*F*	Sig.
Corrected model	1153.781[a]	12	96.148	13.686	.000
Intercept	7700.717	1	7700.717	1096.162	.000
Delay	97.883	5	19.577	2.787	.026
Jitter	931.034	4	232.759	33.132	.000
Drop	46.939	3	15.646	2.227	.096
Error	372.334	53	7.025		
Total	20843.989	66			
Corrected total	1526.115	65			

[a] *R* squared = .756 (Adjusted *R* squared = .701).

Table 9
List of experiment videos

Exp #	Delay (ms)	Jitter (ms)	Drop (%)	PSNR (dB)
e0r	0	0	0	Original
e2r	100	0	0	41
e3r	200	0	0	35
e4r	300	0	0	31
e26	0	0	5	28
e28	400	0	5	26
e29	0	2	5	24
e20	400	2	0	22
e30	50	2	5	21
e19	300	2	0	19
e12r	100	5	0	18
e32r	0	5	5	17
e8	200	10	0	16
e11	50	5	0	15
e15	400	5	0	14

experience before the experiments. Once the experiment procedure is explained by the researchers, they were asked to watch 15 video sequences (see Table 9) all generated from the same video source. Viewing conditions were controlled by using the same laptop machine for all the experiments. After viewing each video, the subjects were asked to provide a quality score for the video shown and a clinical decision making capability score within the context of a medical consultation. The data collection sheet used during the experiments is provided in Appendix A.

6. Phase 2: results and data analysis

Initial analysis was focused on the sample characteristics. As reported before, the subjects were asked to fill out a simple questionnaire before the experiments. Results of this survey indicate that out of 7 IST doctoral students none of them had any telemedicine experience before the experiment. On the other hand, 2 out of 8 optometrists (OD) had been involved in few telemedicine cases. Those who experienced telemedicine before were asked to describe their experience shortly. One reported the experience as "sent retinal image of possible urenal melanoma to retinal specialist for second opinion", and this was repeated by the subject at three different instances. The other subject reported the experience as a brief demonstration of telemedicine in class. Subjects who have extensive telemedicine experience might perceive the experiments differently. However, since none of these subjects had extensive experiences with direct telediagnosis, we have decided to include these subjects in the experiments with others who have no telemedicine experience.

The next step in the analysis was to identify the mean and standard deviation values of the quality and decision making capability assessments. The scale used for quality (continuous scale presented in Appendix A) is evaluated as a 100 point scale. Descriptive analysis of the quality score for each video sequence by all, only IST, and only OD subjects are presented in Table 10. Quality scores (QS) which were on a 0–100 scale were later on converted into mean opinion score (MOS). MOS is a 5 point scale where 1 corresponds to "poor" and 5 corresponds to "excellent". It divides 0–100 scale into 5 equal sections.

The results show that even the quality of the original video, where no degradation was introduced, does not reach the excellent score. This is caused by the original

Table 10
Descriptive statistics of quality scores

		All Subj. (N=15)		IST Subj. (N=7)		OD Subj (N=8)	
Exp #	PSNR	Mean	SD	Mean	SD	Mean	SD
e15	14	18.93	12.898	18.43	14.864	19.38	11.952
e11	15	26.87	17.639	31.29	21.101	23.00	14.283
e8	16	12.33	11.697	16.14	15.604	9.00	6.141
e32r	17	35.20	16.258	37.71	13.829	33.00	18.784
e12r	18	26.33	14.351	31.43	17.396	21.88	10.190
e19	19	34.53	14.904	32.71	16.059	36.13	14.730
e30	21	39.00	17.744	43.00	17.963	35.50	17.976
e20	22	33.80	19.908	45.43	18.937	23.63	15.222
e29	24	33.87	18.524	35.14	21.287	32.75	17.169
e28	26	47.07	14.434	50.43	15.841	44.13	13.432
e26	28	45.33	18.469	51.29	14.648	40.13	20.781
e4r	31	49.80	16.524	50.29	12.162	49.38	20.466
e3r	35	52.40	14.870	59.57	10.470	46.13	15.869
e2r	41	50.27	16.594	54.57	18.814	46.50	14.580
e0r	Orig.	47.07	17.310	52.86	15.214	42.00	18.385

Table 11
Descriptive statistics of capability scores

		All Subj. (N=15)		IST Subj. (N=7)		OD Subj (N=8)	
Exp #	PSNR	Mean	SD	Mean	SD	Mean	SD
e15	14	1.60	.737	1.43	.787	1.75	.707
e11	15	2.27	1.100	2.29	1.113	2.25	1.165
e8	16	1.33	.617	1.57	.787	1.13	.354
e32r	17	2.40	.910	2.29	.951	2.50	.926
e12r	18	2.20	1.014	2.29	1.380	2.13	.641
e19	19	2.73	.961	2.57	.976	2.88	.991
e30	21	2.80	1.146	3.00	1.000	2.62	1.302
e20	22	2.73	1.163	3.00	.816	2.50	1.414
e29	24	2.47	1.060	2.43	1.272	2.50	.926
e28	26	3.47	.834	3.57	.976	3.38	.744
e26	28	3.33	.976	3.71	.488	3.00	1.195
e4r	31	3.47	.834	3.57	.787	3.38	.916
e3r	35	3.33	.976	3.57	.787	3.13	1.126
e2r	41	3.33	.816	3.43	.976	3.25	.707
e0r	Orig.	3.20	.862	3.43	.535	3.00	1.069

Table 12
Frequency analysis of capability evaluations ($N=15$)

Exp #	PSNR	1 (%)	2 (%)	3 (%)	4 (%)	5 (%)
e15	14	53.3	33.3	13.3	0.0	0.0
e11	15	33.3	20.0	33.3	13.3	0.0
e8	16	73.3	20.0	6.7	0.0	0.0
e32r	17	13.3	46.7	36.7	13.3	0.0
e12r	18	20.0	53.3	20.0	0.0	6.7
e19	19	13.3	20.0	46.7	20.0	0.0
e30	21	13.3	33.3	13.3	40.0	0.0
e20	22	20.0	20.0	26.7	33.3	0.0
e29	24	20.0	33.3	26.7	20.0	0.0
e28	26	0.0	13.3	33.3	46.7	6.7
e26	28	6.7	13.3	20.0	60.0	0.0
e4r	31	0.0	13.3	33.3	46.7	6.7
e3r	35	6.7	13.3	20.0	60.0	0.0
e2r	41	0.0	20.0	26.7	53.3	0.0
e0r	Orig.	6.7	6.7	46.7	40.0	0.0

recording of the video clip where the camera was zoomed into a patient's eye in order to make the examination. During the zoom effect, the camera had a hard time focusing on the eye since the contractions in the eye caused by rapid blinking were very fast. This introduced a blurry effect in the original video which cannot be eliminated. Therefore, the quality scores for the original video were below excellent.

The clinical decision making capability scale (presented in Appendix A) is evaluated as a 5 point scale where "1" corresponds to "I cannot make a clinical decision" and "5" corresponds to "I can easily make a medical decision". This scale is developed to capture the relationship between the quality perceptions and how it translates into clinical decision making capability. Descriptive analysis of the capability score (CS) for each video sequence by all, only IST, and only OD subjects are presented in Table 11.

Frequency analysis of the capability evaluations are presented in Table 12. More than half of the subjects rated 6 of the videos in 1–2 capability score zone, and the remaining 9 videos in 3–4 capability score zone. Five of the six videos also had the lowest objective quality scores (PSNR < 20 db). Videos with high objective quality scores (PSNR > 25 db) also received high capability scores by the majority of the subjects (80% or more) and at least half of the subjects rated these videos in the 4 capability score. These results imply that when the objective PSNR score is higher than 25 db, the subjects felt reasonably comfortable to make a decision. One of the expert subjects assigned the highest confidence score to three of the videos. Overall, the scores assigned by this subject were higher (3 or 4) compared to the scores assigned by the other subjects. These scores indicate that based on the information provided in these videos, the subject felt comfortable enough to make a decision if this was an emergency case.

Table 13
Correlation table for e11 (PSNR = 15 db)

	e11Order	e1QSs	e11CS	e11MOS
e11Order	1.000			
e11QS	−.513	1.000		
e11CS	−.626[a]	.819[b]	1.000	
e11MOS	−.338	.948[b]	.755[b]	1.000

[a] Correlation is significant at the 0.05 level.
[b] Correlation is significant at the 0.01 level.

Table 14
Correlation table for e20 (PSNR=22 db)

	e20Order	e20QS	e20CS	e20MOS
e20Order	1.000			
e20QS	−.072	1.000		
e20CS	−.023	.763[a]	1.000	
e20MOS	−.036	.973[a]	.726[a]	1.000

[a] Correlation is significant at the 0.01 level.

After the descriptive analysis, difference between the quality and decision making capability perception of two subject groups (IST and OD) was analyzed using One-Way ANOVA. The results of the analysis indicate that there is no significant difference between the expert groups in terms of their quality and capability perceptions. Therefore, the analysis conducted does not differentiate between these two groups.

To better understand the relationship between quality and capability values, we calculated Pearson correlations between Order (the order video was presented during subjective tests), QS(1–100 quality score assigned during subjective tests), CS(1–5 capability score assigned during subjective tests), and MOS(1–5 quality score that was driven from the 1–100 quality scores assigned during subjective tests). The order of video sequences presented to each subject was randomized during subjective experiments, because the order in which the video sequence was presented to the subject might have an effect on the final perceived quality. Therefore, Order was also included in the correlations as a fourth variable to understand if there is any correlation between the order of presentation and the perceived scores assigned by the subjects.

Table 15
Correlation table for e26 (PSNR = 28 db)

	e26Order	e26QS	e26CS	e26MOS
e26Order	1.000			
e26QS	.734[a]	1.000		
e26CS	.396	.734[a]	1.000	
e26MOS	.653[a]	.974[a]	.725[a]	1.000

[a] Correlation is significant at the 0.01 level.

Table 16
Correlation table for e29 (PSNR=24 db)

	e29Order	e29QS	e29CS	e29MOS
e29Order	1.000			
e29QS	.174	1.000		
e29CS	.037	.640[a]	1.000	
e29MOS	.088	.963[b]	.722[b]	1.000

[a] Correlation is significant at the 0.05 level.
[b] Correlation is significant at the 0.01 level.

Table 17
Correlation table for e2r (PSNR=41 db)

	e2rOrder	e2rQS	e2rCS	e2rMOS
e2rOrder	1.000			
e2rQS	.807[b]	1.000		
e2rCS	.616[a]	.694[b]	1.000	
e2rMOS	.802[b]	.944[b]	.660[b]	1.000

[a] Correlation is significant at the 0.05 level.
[b] Correlation is significant at the 0.01 level.

Table 18
Correlation table for e30 (PSNR=21 db)

	e30Order	e30QS	e30CS	e30MOS
e30Order	1.000			
e30QS	−.456	1.000		
e30CS	−.641[a]	.839[b]	1.000	
e30MOS	−.404	.953[b]	.766[b]	1.000

[a] Correlation is significant at the 0.05 level.
[b] Correlation is significant at the 0.01 level.

Table 19
Correlation table for e32r (PSNR=17 db)

	e32rOrder	e32rQS	e32rCS	e32MOS
e32rOrder	1.000			
e32rQS	−.260	1.000		
e32rCS	−.017	.795[a]	1.000	
e32MOS	−.149	.945[a]	.734[a]	1.000

[a] Correlation is significant at the 0.01 level.

Table 20
Correlation table for e4r (PSNR=31 db)

	e4rOrder	e4rQS	e4rCS	e4rMOS
e4rOrder	1.000			
e4rQS	.150	1.000		
e4rCS	.065	.686[a]	1.000	
e4rMOS	.164	.953[a]	.724[a]	1.000

[a] Correlation is significant at the 0.01 level.

Table 21
Correlation table for e8 (PSNR=16 db)

	e8Order	e8QS	e8CS	e8MOS
e8Order	1.000			
e8QS	−.240	1.000		
e8CS	−.172	.775[a]	1.000	
e8MOS	−.226	.871[a]	.715[a]	1.000

[a] Correlation is significant at the 0.01 level.

Table 22
Correlation table for e12r (PSNR=18 db)

	e12rOrder	e12rQS	e12rCS	e12rMOS
e12rOrder	1.000			
e12rQS	0.348	1.000		
e12rCS	−0.068	.653[a]	1.000	
e12rMOS	0.276	.908[a]	.681[a]	1.000

[a] Correlation is significant at the 0.01 level.

Table 23
Correlation table for e19 (PSNR=19 db)

	e19Order	e19QS	e19CS	e19MOS
e19Order	1.000			
e19QS	−.210	1.000		
e19CS	−.197	.769[b]	1.000	
e19MOS	−.429	.939[b]	.641[a]	1.000

[a] Correlation is significant at the 0.05 level.
[b] Correlation is significant at the 0.01 level.

Table 24
Correlation table for e28 (PSNR=26 db)

	e28Order	e28QS	e28CS	e28MOS
e28Order	1.000			
e28QS	.253	1.000		
e28CS	.408	.561[a]	1.000	
e28MOS	.237	.929[b]	.560[a]	1.000

[a] Correlation is significant at the 0.05 level.
[b] Correlation is significant at the 0.01 level.

Table 25
Correlation table for e3r (PSNR=35 db)

	e3rOrder	e3rQS	e3rCS	e3rMOS
e3rOrder	1.000			
e3rQS	−.313	1.000		
e3rCS	−.253	.753[b]	1.000	
e3rMOS	−.248	.940[b]	.580[a]	1.000

[a] Correlation is significant at the 0.05 level.
[b] Correlation is significant at the 0.01 level.

Table 26
Correlation table for original sequence (e0)

	e0rOrder	e0rQS	e0rCS	e0rMOS
e0rOrder	1.000			
e0rQS	0.435	1.000		
e0rCS	0.467	0.507	1.000	
e0rMOS	0.442	.939[a]	0.493	1.000

[a] Correlation is significant at the 0.01 level.

It was expected that the quality score and the capability score will significantly correlate. Similarly, MOS, which is derived from the quality score, was expected to correlate with the capability score. If these assumptions were correct, the decision making capability of a medical professional could have been predicted using the perceived quality scores. The results of the Pearson's correlation are illustrated in Tables 13–27. As shown in Tables 13–22, for 10 out of 15 video sequences utilized during the subjective tests, the MOS values were significantly correlated with the CS values at 0.01 level. MOS and CS for other three video sequences were correlated but at a lower significance level (0.05) as illustrated in Tables 23–25. As shown in Tables 26–27, the correlation between the two scores was not statistically significant for two of the video sequences (one being the non-impaired video — e0r). However, one of these videos (e15) showed a correlation between the QS and CS values (significant at 0.01 level).

7. Discussion

Results presented in the previous section support the initial hypothesis that the quality score does not necessarily correspond with a medical decision making capability score. Study subjects in their written and verbal comments emphasized the importance of the critical frames that they will base their decision on. They sometimes were able to make decisions with an overall low quality video sequence since the impairment in that sequence occurred at non-critical frames. They also mentioned that if they were able to see that critical frame, they will not need to watch the rest of the video. This finding has important implications on videoconferencing applications that are built to support telemedicine. In a telemedicine environment, it is important for the medical provider to have control on the quality. If the application provides functionality to increase the quality when there is request from the user, the decision making capability will be positively affected.

Table 27
Correlation table for e15 (PSNR = 14 db)

	e15Order	e15QS	e15CS	e15MOS
e15Order	1.000			
e15QS	−0.245	1.000		
e15CS	−0.502	.651[a]	1.000	
e15MOS	−0.090	.875[a]	0.496	1.000

[a] Correlation is significant at the 0.01 level.

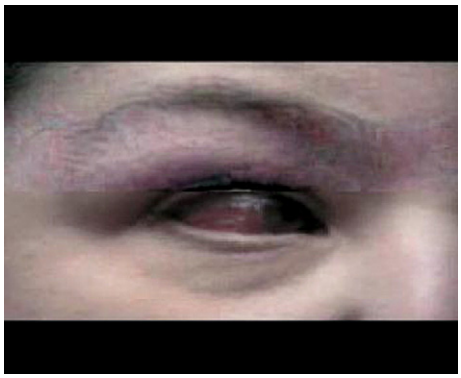


Fig. 5. Experiment 12r critical frame.

An interesting case occurred when one subject evaluated experiment 12r (e12r) with a 5 CS which indicates that the subject felt that a medical decision can be made easily based on this video. If we check in Table 9, the PSNR value calculated for this video sequence was 18dB, which indicates bad quality compared to the unimpaired video. Figs. 5 and 6 present two different frames from the same video.

Fig. 5 represents the case where the impairments happening on the network level did not affect the critical frame which is useful for the medical expert in making a decision. However, as Fig. 6 illustrates, the overall quality of the video sequence was quite poor.

8. Conclusion

This study investigated the relationship between objective and subjective quality measures, as well as

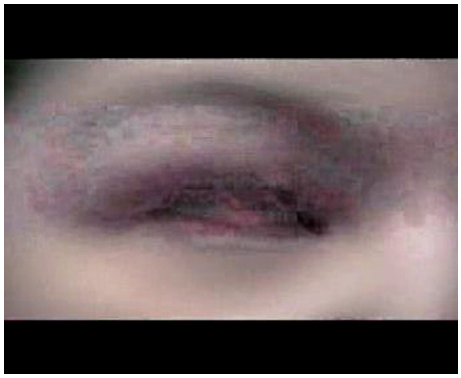


Fig. 6. Experiment 12r 10 frames before the critical frame.

clinical decision making capability that is affected by the impairments that occur during the transmission of video over the Internet. It was one of the few studies that emphasized differences in quality perception when the video and application is in the medical context and a medical expert is the user depending on the video being transmitted over an Internet to make a decision.

In this first phase, effects of network impairments on video quality for a telemedicine video were studied. Contributions are as follows: (1) significant effect of jitter on video with high movement is identified, (2) cutoff points for the three values were provided to achieve reasonable video quality. The next phase conducted a subjective quality measurement and investigated the correlations between objective, subjective measurements and decision making capability.

Many studies concluded that telemedicine improves the access to care. The goal of this study is to improve access to telemedicine by understanding the challenges for Internet-based telemedicine for ophthalmology domain. By repeating the same experiments for other domains, it is possible to create a quality metric for Internet-based telemedicine. Existence of quality metrics can enable the development of smart tools that can handle challenges introduced by the Internet with minimum involvement from the users. Stakeholders of Internet-based telemedicine have unique challenges to face. For doctors, the biggest challenge is to rely on a video tool running over the Internet and its quality. Can they consult or diagnose? What happens when there is congestion on the Internet and the quality of the images or video degrades to an unacceptable level? When should the doctors give up?

The first phase of this study aimed to take a step at understanding the effects of network impairments on a specific telemedicine video (general eye examination). An objective quality database was built and certain thresholds were defined for the experimental testbed. Findings indicate that jitter effects are significant on degradation of video quality and video tools need to provide solutions for handling this parameter in order to achieve successful telemedicine implementations over the Internet.

The second phase of this study built the linkage between objective quality measurements calculated using mathematical methods and perceived quality evaluated by human subjects. This phase of the study identified that the viewing order of sequences does not correlate with the quality or capability scores other than two sequences (e2r and e26). Moreover, objective quality scores (PSNR) are most of the time significantly correlated (0.05 or 0.01 level) with subjective quality and capability scores. However, for some sequences the correlation between objective and subjective scores is weaker or does not exist because (1) the position of degradation is important, and (2) the quality of the critical frames affect decision making capability.

These findings are useful while developing algorithms for application level QoS and decision support tools. The findings of this work suggest that further studies that measure quality of video using sequences from real-world telemedicine examples is necessary to understand the effects of quality on medical decision making. Only then one can improve the existing applications to serve the needs of medical professionals.

There are some limitations to this study. First of all, quality of the original video sequence was not perfect. Hence, it is not easy to identify the reasons for the degradation in the video quality once it is transmitted over the Internet. This is one of the reasons for the original video not scoring high quality values from the subjects of this study. They were complaining more about the blurry effect than the pixelization and smudging effects. Second, the study results are highly dependent on the application area used in this video. It was for general eye examination and the results of this study cannot be generalized for other domains such as telemental health before further research is completed. This limitation is driven by the large variety of specialty areas that exist in telemedicine. Each of these specialty areas has its own information requirements and decision making capability levels. Therefore, this study provides an important step in initiating the creation of a quality metric which can be developed through a series of studies that focus on different specialty areas. Third, the sample size for the subjective tests is low compared to other studies in the field of Information Systems. In order to improve the generalizability of the results, the same experiment can be applied to different sample groups in future studies. Authors are willing to distribute the degraded video sequences to those who would like to apply subjective quality tests. One other limitation we acknowledge is the use of one video sequence from the real-world telemedicine. This is highly due to HIPAA and privacy regulations in place in the U.S. which limits researcher's ability to procure more videos.

Besides these limitations, this study provides insights from a measurement-research specific to telemedicine videos. It is important to study this domain in isolation from others to get better understanding of the user needs which will eventually increase the opportunities around the world to receive medical care via telemedicine.

As for the future research, further experiments will be conducted by adding other video codecs as a new factor. Video codec implementation and the application used to transmit video may also have significant effects on the

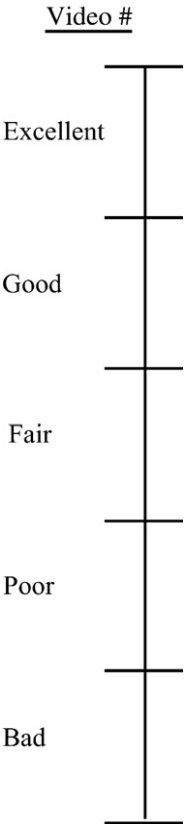
video quality. Future work will identify the resistance of other codecs to network impairments. This study provides a structured outline for conducting quality assessment studies in telemedicine context. Similar experiments can be conducted under different medical domains and results can be compared. Utilizing these results, a large quality database can be developed to support clinical decision making at the point of service. Currently, we are working on building an advanced Internet videoconferencing tool that incorporates the results of this research.

Acknowledgements

We would like to thank the Regenstrief Institute, Inc. and the National Library of Medicine for sharing the video sequence utilized in this study with us. We specifically would like to thank Dr. Michael Weiner and Dr. Craig Locatis for helping us with the arrangements.

Appendix A. Subjective quality evaluation sheet sample

Evaluation Sheet — Subject #



Assuming that you were asked to handle the case of this patient based on this video, which one of the following best represents your opinion?

___ I can easily make a medical decision.
___ I can make a medical decision, up to a reasonable certainty.
___ I can make a clinical decision only if this is an emergency case.
___ I would rather not make a clinical decision.
___ I cannot make a clinical decision.

Please feel free to provide any comments regarding the video quality or the quality scale:

_____
_____
_____
_____

References

[1] R. Blum, Network Performance Open Source Tool Kit: Using Netperf, tcptrace, NIST Net, and SSFNet, Wiley Publishing Inc., Indianapolis, Indiana, 2003.
[2] M. Carson, D. Santay, NIST Net: a Linux-based network emulation tool, ACM SIGCOMM Computer Communication Review 33 (2003) 111–126.
[3] J.R. Coleman, HMOs and the future of telemedicine and telehealth part 1, The Case Manager 13 (2002) 36–40.
[4] P.C. Cosman, R.M. Gray, R.A. Olshen, Evaluating quality of compressed medical images: SNR, subjective rating, and diagnostic accuracy, Proceedings of the IEEE 82 (1994) 919–932.
[5] P. Dev, D. Harris, D. Gutierrez, A. Shah, S. Senger, End-to-end performance measurement of Internet based medical applications, presented at American Medical Informatics Association (AMIA) Symposium, 2002.
[6] P. Dev, K. Montgomery, S. Senger, W.L. Heinrichs, S. Srivastava, K. Waldron, Simulated medical learning environments on the Internet, Journal of the American Medical Informatics Association 9 (2002) 437–447.
[7] R.H. Eikelboom, K. Yogesan, C.J. Barry, I.J. Constable, L. Jitskaia, P.H. House, M.L. Tay-Kearney, Methods and limits of digital image compression of retinal images for telemedicine, Investigative Ophthalmology and Visual Science 41 (2000) 1916–1924.
[8] K. Fall, Network emulation in the Vint/NS simulator, presented at Fourth IEEE Symposium on Computers and Communications, Sharm El Sheik, Red Sea, Egypt, 1999.
[9] C.W. Flowers, Technology guidelines by type of clinical service: ophthalmology, in: J. Tracy (Ed.), Telehealth Technology Guidelines, Office for the Advancement of Telehealth (OAT), Health Resources and Services Administration, 1999, http://telehealth.hrsa.gov/pubs/tech/techhome.htm, accessed on May 24, 2005.
[10] M. Hassan, A. Nayandoro, M. Atiquzzaman, Internet telephony: services, technical challenges, and products, IEEE Communications Magazine 38 (2000) 96–103.
[11] Sun Microsystems Inc. Documentation: JMStudio User's Guide, accessed January 16, 2007, [available at http://java.sun.com/products/java-media/jmf/2.1.1/jmstudio/jmstudio.html].

[12] R. Jain, The Art of Computer Systems Performance Analysis: Techniques for Experimental Design, Measurement, Simulation, and Modeling, John Wiley & Sons, Inc., New York, NY, 1991.
[13] A.P. Markopoulou, Assessing the Quality of Multimedia Communications Over Internet Backbone Networks, in Department of Electrical Engineering, Stanford University, Stanford, CA, 2002.
[14] A.P. Markopoulou, F.A. Tobagi, M.J. Karam, Assessing the quality of voice communications over Internet backbones, IEEE/ACM Transactions on Networking 11 (2003) 747–760.
[15] S. Mohamed, Automatic Evaluation of Real-Time Multimedia Quality: a Neural Network Approach, University of Rennes I, Rennes, 2003.
[16] M.H. Pinson, S. Wolf, A new standardized method for objectively measuring video quality, IEEE Transactions on Broadcasting 50 (2004) 312–322.
[17] A. Przelaskowski, Vector quality measure of lossy compressed medical images, Computers in Biology and Medicine 34 (2004) 193–207.
[18] D.A. Rosenthal, Analyses of selected variables effecting video streamed over IP, International Journal of Network Management 14 (2004) 193–211.
[19] H. Schulzrinne, S. Casner, R. Frederick, V. Jacobson, RTP: A transport protocol for real-time applications, Internet Engineering Task Force (IETF), vol. 3550, RFC, July 2003.
[20] P. Taylor, Evaluating telemedicine systems and services, Journal of Telemedicine and Telecare 11 (2005) 167–177.
[21] A. Webster, C. Jones, M. Pinson, S. Voran, S. Wolf, An objective video quality assessment system based on human perception, presented at Storage and Retrieval for Image and Video Databases — Human Vision, Visual Processing and Digital Display TV, San Jose, CA, USA, 1993.
[22] E.E. Westberg, R.A. Miller, The basis for using the internet to support the information needs of primary care, Journal of the American Medical Informatics Association 6 (1999) 6–25.
[23] S. Wolf, M. Pinson, Video quality measurement techniques, National Telecommunication and Information Administration NTIA Report 02-392, U.S. DEPARTMENT OF COMMERCE, June 2002.

Bengisu Tulu, Ph.D., is an Assistant Professor in the Department of Management at Worcester Polytechnic Institute, Massachusetts. She is also a research fellow in the Kay Center for E-Health Research at Claremont Graduate University. Se holds a B.S. in Mathematics from Middle East Technical University, Turkey and an M.S and Ph.D. from the School of Information Systems and Technology, Claremont Graduate University, California. Her research is focused on the use of information technology within healthcare. Her current research projects include user centric personal health records for people with disabilities, Internet-based telemedicine, business value of telemedicine and healthcare outsourcing. She has published in respected scholarly journals and refereed conferences including Telemedicine & e-Health Journal, International Journal of Healthcare Technology & Management, IEEE Network, IEEE J. on Selected Areas in Communications, Communications of AIS, etc.

Samir Chatterjee is a Professor in the School of Information Systems & Technology and Founding Director of the Network Convergence Laboratory at Claremont Graduate University, California. Prior to that, he taught at the CIS department of J Mack Robinson College of Business, Georgia State University, in Atlanta. He holds a B.E (Hons.) in Electronics Engineering from Jadavpur University, India and an M.S and Ph.D. from the School of Computer Science, University of Central Florida. He is widely recognized as an expert in the areas of Next-Generation Networking, Voice and Video over IP, and Network Security. His current research interests include the design of secured IT-based systems to be used in application fields such as healthcare information systems, P2P computing, ad hoc collaboration, creativity and telemedicine. He has published over 75 articles in respected scholarly journals and refereed conferences including IEEE Network, IEEE J. on Selected Areas in Communications, Communications of the ACM, Computer Networks, International Journal of Healthcare Technology & Management, Telemedicine & e-Health Journal, Information Systems Frontiers, Computer Communication, IEEE IT Professional, ACM CCR, Communications of AIS, Journal of Internet Technology etc. He has actively contributed towards designing middleware for multimedia within Internet2 which led to the establishment of the ITU-T standard called H.350. He is principal investigator on several NSF grants and has received funding from numerous private corporations for his research. He is on the editorial board of IJBDN and JITTA. He is the founding Program Chair for the International Conference on Design Science Research in IS&T (DESRIST 2006, 2007). He is Vice Chair of EntNet Technical Committee for IEEE Communications Society and serves on the TPC for WITS 2007, ICIS 2006, IEEE Healthcom 2006, IEEE MASS'05 and Workshop Chair at EntNet@Supercom 2005. He has been an entrepreneur and successfully co-founded a startup company VoiceCore Technologies Inc. in 2000.