Real-time Video Quality Assessment in Packet Networks: A Neural Network Model

Samir Mohamed, Gerardo Rubino
IRISA/INRIA, Campus du Beaulieu
35042 Rennes, France[1]

Hossam Afifi
INT, Evry
France

Francisco Cervantes
ITAM, Mexico City
Mexico

Abstract – *There is a great demand to assess video quality transmitted in real time over packet networks, and to make this assessment in real time too. Quality assessment is achieved using two types of methods: objective or subjective. Subjective methods give more reliable results (objective ones do not correlate well with human perception), but unfortunately, they are not suitable to real-time applications and are difficult to carry out. In this paper, we show how Artificial Neural Networks (ANN) can be used to mimic the way by which a group of human subjects assess video quality when this video is distorted by certain quality-affecting parameters (e.g. packet loss rate, loss distribution, bit rate, frame rate, encoded frame type, etc.). Our method can be used to measure in real time the subjective video quality with very good precision. In order to illustrate its applicability, we chose to assess the quality of video flows transmitted over IP networks and we carried out subjective quality tests for video distorted by variations of those parameters.*

Keywords: Packet video, Neural Networks, Real-time video transmission, Video quality assessment.

1. Introduction

Many real-time video transmission applications over the Internet have appeared in the last few years. We find today: video phones, video conferencing, video streaming, tele-medical applications, distance learning, telepresence, video on demand, etc., with a different number of requirements in bandwidth and perceived quality. This gives rise to a need for assessing the quality of the transmitted video in real time.

Quality in encoders is also a way to "fit" the stream into the available global channel bandwidth. In video codecs using temporal compression (ex: MPEG, H.261, H.263…), a quality factor parameter is usually used to reduce the output stream bandwidth and to reduce, in the same time, the assessed quality (yet before any transmission).

Now, if we consider the parameters that affect the quality (*quality-affecting* parameters) of video transmission over packet networks, we can classify them as follows:

- Coding and compression parameters: They control the amount of quality losses that happen during the encoding process; so they depend on the type of the encoding algorithm (MPEG, H26x, etc.), the output bit rate, the frame rate, the temporal relation among frame kinds (I, B, P, etc.), etc. [5] [22];
- Network parameters: They result from the packetization of the video stream [17] and the transmission in real-time, such as the packet loss rate, the loss distribution, the delay, the delay variation (jitter), etc. [6][1][2];
- Other parameters like the nature of the scene (e.g. amount of motion, color, contrast, image size, etc.) [5][21][20].

Since in this paper we concentrate only on pure video applications, we do not take into account parameters such as lip synchronization or other audio aspects.

It is clear that quality is not linearly proportional to the variation of these parameters. The determination of the quality is a very complex problem, and there is no mathematical model that can take into account the effects of all these parameters.

There are two approaches to assess the quality: objective and subjective methods. The objective methods [23] measure the quality based on mathematical analysis that compare original and distorted video sequences. Some existing methods are MSE (Mean Square Error) or PSNR (Peak Signal to Noise Ratio) which measures the quality by a simple difference between frames. There are other methods that are much more complicated like the moving

[1] This work was partially supported by the European ITEA Project 99011 "RTIPA" (Real-Time Internet Platform Architectures).

picture quality metric (MPQM), and the normalized video fidelity metric (NVFM) [22].

On the other hand, subjective methods [16] measure the overall perceived video quality. They are carried out by human subjects. The most commonly used for video quality evaluation is the Mean Opinion Score (MOS), recommended by the ITU. It consists in having ***n*** subjects viewing the distorted video sequences in order to rate their quality, according to a predefined quality scale. That is, human subjects are trained to "build" a mapping between the quality scale and a set of processed video sequences.

Although MOS studies have served as the basis for analyzing many aspects of signal processing, they present several limitations: a) very stringent environments are required; b) the process can not be automated; c) it is very costly and time consuming to repeat it frequently. Consequently, it is impossible to use it in real-time quality assessment. On the other hand, the disadvantages of objective methods are: a) they do not correlate well with human visual perception[2]; b) they require high calculation power, and are time consuming (they usually operate at the pixel level); c) it is very hard to adapt them to real-time quality assessment, as they work on both the original video sequence and the transmitted/distorted one. Then, instead of looking for algorithms to objectively measure video quality, why do not we build a hybrid system that takes into consideration subjective measurements, and which behavior is close to that of humans when they evaluate video quality?

In this paper, we address this question by describing a method for developing such an automaton. Our problem has two aspects: first, a classification one, to map the non-linear relation between the parameters and the quality; second, a prediction one, to evaluate the quality as a function of the quality-affecting parameters in an operational environment.

We believe that artificial neural networks (ANN) are an appropriate tool to solve this two-fold problem [18]. We illustrate our approach by building a system that takes advantage of the benefits offered by ANN to capture the nonlinear mapping between several non-subjective measures (i.e. the quality-affecting parameters) of video sequences transmitted over packet switched networks and the quality scale carried out by a group of humans subjects during an MOS experiment. More details on our proposal are given in [13].

The structure of this paper is as follows. The next Section situates our study in its context, by describing related works. Section 3 presents our proposal in detail. Our results are the object of Section 5. The last Section presents our conclusions and future research directions.

2. Related Works

In previous work, we showed how to use ANN to measure in real-time audio quality when this audio is transmitted over packet networks [11]. Based on this technique, we developed a new control mechanism that permits a better use of bandwidth and the delivery of the best possible audio quality given the current network situation [12].

The work in [14] presents a methodology for video quality assessment using objective parameters based on image segmentation. An image encoded by MPEG-2 is segmented into three regions: plane, edges, and texture; then, a set of objective parameters is assigned to each region. After that, a perception-based model is defined by computing the relationship between objective measures and results of subjective tests.

In [2], the authors study the effect of both loss and jitter on the perceptual quality of video. They argue that, if there is no mechanism to mask the effect of jitter, the perceived quality degrades in the same way as it degrades with losses. In [6] [21] [20] [8], the authors analyze the effect of audio synchronization on the perceived video quality; they quantify the benefits of audio synchronization on the overall quality of the flow.

The main goal of [5] is to study the effect of the frame rate for different standard video sequences on the overall perceived quality. The work presented in [1] is mainly a study of the packet loss effects on MPEG video streams. The authors show also the effect of loss rate on the different types of MPEG frames. While in [22], a study of the effect of the bit rate on the objective quality metrics (PSNR, NVFM, and MPQM) is presented. The effect of the number of consecutively lost packets on the video quality is analyzed in [6].

In [9], the authors present how to use ANN to predict packet loss during a real-time video transmission over a packet network as a function of the inter-packet delay variation. ANN are used also in video compression with compression ratio that goes from 500:1 to 1000:1 for moving gray-scale images

[2] By the way, this claim comes from comparing their results to those obtained from subjective methods.

and full-color video sequences respectively [3]. Furthermore, they are used in a variety of image processing techniques which go from image enlargement and fusion to image segmentation [4].

3. Description of our Method

In this Section, we describe the overall steps that should be followed in order to build a tool to automatically assess in real time the subjective quality of real-time video transmitted over packet networks. The aim of this method is to use ANN to model and evaluate in real time how human subjects estimate video quality when distorted by changes in the quality-affecting parameters.

We start by defining a set of static information that will affect the general quality perception. We must choose the most effective quality-affecting parameters corresponding to the type of video application and to the network supporting the transmission (see Section 5.2).

Once the quality-affecting parameters are identified, for each one we should find the two extremes and the most frequent occurrences of its values. This can be done either by real measurement or by using simulation techniques. For example, if the percentage loss rate is expected to vary from 0 to 10 %, then we may use 0, 1, 2, 5, and 10 % as the typical values for the loss rate parameter. If we call configuration of the set of quality-affecting parameters a set of values for each one, the total number of possible configurations is usually large. We must then select a part of this large cardinality set, which will be used as the input data of the ANN in the learning phase.

Depending on the transmission configuration, a simulation environment or a testbed should be implemented. This environment is used to send video sequences from the source to the destination and to control the underlying packet network. Every configuration in the defined input data must be mapped into the system composed of the network, the source and the receiver. For example, in IP networks, the source controls bit rate, the frame rate and the encoding algorithm, and it sends RTP video packets; the router controls the loss rate, the loss distribution, the delay and the jitter; the destination stores the transmitted video sequence and collects the corresponding values of the parameters. Of course, one can generate the distorted signal by simulation. Then, by operating the testbed or the artificial simulation, we produce and store a set of distorted signals along with their corresponding values of the parameters.

After completing the video database, a subjective quality test should be carried out. There are several subjective quality methods in the recommendations of the ITU-R [15] (see Section 4). The video database should be shuffled in such a way to avoid the effect of the last sequence on the judgment of the current one. A group of human subjects is then invited to evaluate the quality of the distorted video sequences (i.e. every subject gives each video sequence a score from the predefined quality scale).

The next step is to calculate the MOS values for all the video sequences. Based on the results obtained by the human subjects, a prescreening and statistical analysis may be carried out to remove the grading of the individuals suspected to give unreliable results [15]. After that, we store the MOS values and the corresponding parameters' values in another database (Quality Database).

Then, a suitable neural network architecture and a training algorithm should be selected. We chose, as in other applications fields, a three-layered feedforward network and the backpropagation training. The quality database is divided into two parts: one to train the ANN and the other to test its accuracy. The trained ANN will emulate the subjective quality measure for any given values of the parameters (not necessary among the training database). The overall procedure can be repeated, if necessary, to improve the ANN.

Once a stable neural network configuration is obtained, the ANN's architecture and weights can be extracted in order to build a concise tool. We decompose such a tool into two parts. The first one collects the values of the quality-affecting parameters. The second part is the trained ANN that will take the given values of the chosen quality-affecting parameters and correspondingly computes the subjective MOS quality score.

1) Operating Mode

Real-time video applications can be considered one-way sessions (i.e. they consist of a sender that produces the video and a receiver that consumes it). This behavior is different from that of audio applications. Indeed, in audio, the interactivity may produce some other parameters (e.g. echo, crosstalk effect, number of participating sources, etc.) that affect the overall quality [11].

In the operating mode when integrated in a video system, our tool will act as shown in Figure 2. In the sender the video source is encoded and affected by

some parameters. Then it is packetized and sent by the transport protocol (e.g. RTP/RTCP) to the receiver. Here again the video quality may be degraded by certain parameters. In the receiver, the flow is decoded and displayed to the end-user.

The interaction between our tool (shaded rectangles in the Figure) and the other elements is as follows. The parameters' collector part probes all the working parameters from the encoder, decoder, packet network and the transport protocol. Then the trained ANN part evaluates video quality as a function of these parameters.

While at the sender side, if necessary, video quality can be sent by the transport protocol from time to time (in RTCP protocol, it can be sent every 5 *sec*.). This means that the frequency update of the parameters and hence the quality evaluation can be done at any time the user wants at the receiver side, while at the sender side the user can have a feedback about the quality at least every 5 *sec.*

4. Subjective Quality Test

To evaluate the quality of video systems (codec, telecommunication, television pictures, etc.), a subjective quality test is used. In this test, a group of human subjects is invited to judge the quality of the video sequence under the system conditions (distortions). There are several recommendations [15][16] that specify strict conditions to be followed in order to carry out the subjective test.

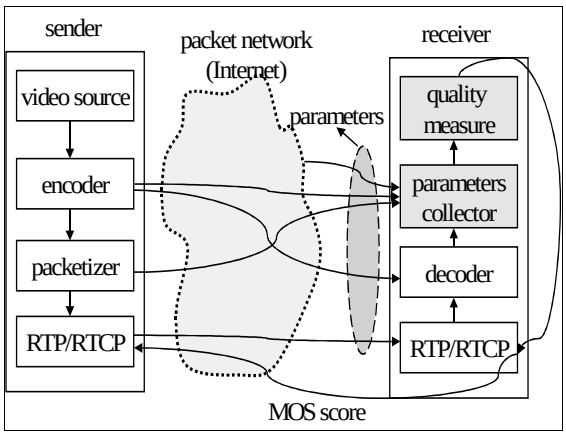

Figure 1. Operation mode for the tool in real-time video system.

1) Subjective quality methods

The main subjective quality methods are Degradation Category Rating (DCR), Pair Comparison (PC) and Absolute Category Rating (ACR). For our study, we are using DCR method, in which a pair of video sequences is presented to each observer, one after the other. They should see the first one, which is not distorted by any impairment, and then the second one, which is the original signal distorted by some configuration of the set of chosen quality-affecting parameters. Figure 2 shows the sequence and timing of presentations for this test. The time values come from the recommendation of the ITU-R [15].

As the observer is faced by two sequences, he/she is asked to assess the overall quality of the distorted sequence with respect to the non-distorted one (reference sequence). Figure 3 depicts the ITU-R nine-grade scale. The observers should give the test presentation a grade from one to nine corresponding to their mental measure of the quality associated with it. It should be noted that there exist several quality scales. We chose this nine-grade one as a tradeoff between precision and dispersion of the subjective evaluations.

Following the ITU-R recommendations, overall subjective tests should be divided into multiple sessions and each session should not last more than 30 minutes. For every session, we should add several dummy sequences (about four or fife). These sequences should not be taken into account in the calculation. Their aim is to be used as training samples for the observers to learn how to give meaningful rates.

5. Results

In order to validate the applicability of our method, we chose to apply it to the assessment of subjective quality of real-time video transmission over IP networks.

1) Simulator Description

To generate the distorted video sequences, we used a tool that encodes a real-time video stream over an IP network into H263 format [7], simulates the packetization of the video stream, decodes the received stream and allows us to handle the simulated lost packets (for instance, for statistical purposes).

The encoder can be parameterized, we can control the bit rate, the frame rate, the intra macro blocs refresh rate (i.e. encode the given macro bloc into intra mode rather than inter mode -this is done to make the stream resistant to losses [10]), image format (QCIF, CIF...), etc. The packetization process is in conformance with RFC 2429 [17].

We used a standard video sequence called *stefan* to test the performance of H26x and MPEG4 codecs. It

contains 300 frames encoded into 30 *frames/s*, and lasts for 10 *sec*. (This follows the ITU recommendations, as usual in the area.) The encoded sequence's format is CIF (352 lines x 288 pixels). The maximum allowed packet length is 536 bytes, in order to avoid the fragmentation of packets between routers.

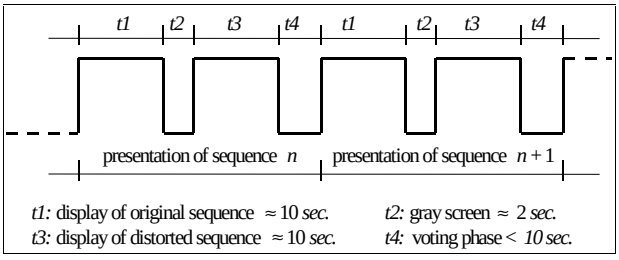

Figure 2. Presentation structure of video sequences in a DCR subjective quality test experiment.

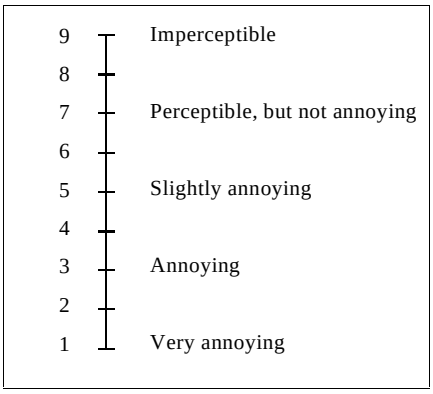

Figure 3. ITU-R nine-grade

2) The quality-affecting parameters

We present here the quality-affecting parameters that we consider having the highest impact on quality:

- The bit rate (BR) in Kbps: this is rate of the actual encoder's output. It is chosen to take four values (256, 512, 768 and 1024 Kbps.). The effect of this parameter on quality is studied in [22].
- The number of frames per second (FR): the original video sequence is encoded at 30 frames per sec. This parameter takes one of 5 values (5, 10, 15 and 30 fps). This is done by the encoder by dropping frames uniformly. A complete study for the effect of this parameter on quality is given in [5].
- The ratio of the encoded intra macro-blocs to inter macro-blocs (G): this is done by the encoder, by changing the refresh rate of the intra macro-blocs in order to make the encoded sequence more or less sensitive to the packet loss [10]. This parameter takes values that vary between 0.053 and 0.4417 depending on the BR and the Fps for the given sequence. We selected five values for it.
- The packet loss rate (LR): the simulator can drop packets randomly and uniformly to satisfy a given percentage loss. This parameter takes five values (0, 1, 2, 4, and 8 %). It is admitted that a loss rate higher than 8 % will drastically reduce the video quality. In the networks where the LR is expected to be higher than this value, some kind of FEC [19] should be used to reduce the effect of losses. There are many studies analyzing the impact of this parameter on quality; see for example [1][5][6].
- The number of consecutively lost packets (CL): we chose to drop packets in bursts of sizes 1 to 5. These values come from real measurements that we performed before [11]. See [6] for a study of the effect of this parameter upon the quality.

The delay and its variation are indirectly considered: they are included in the LR parameter. Indeed, it is known that if a dejittering mechanism with a strict playback buffer length is used, then all the packets arriving after a predefined threshold are considered as lost [2]. So, in this way, all delays and delay variations are mapped into loss.

Given our choices for these parameters' values, we have 4x4x5x5x5 = 2000 different combinations. It is the role of the ANN to interpolate the quality scores for the missed parts of this potentially large input space, learning from the values in the database. We chose to give default values and to compose different combinations by changing only two parameters at a time. This led to 94 combinations.

3) MOS Experiment

The subjective quality test is with conformance to the method Degradation Category Rating (DCR), with a quality scale having 9 points (see Section 4). We divided the test into two sessions, and added 5 distorted sequences to the first session and 4 to the second one. These nine sequences will not be considered in the MOS calculation as their aim is to be used as a training phase for the human subjects. At the same time, they are used to verify how much reliable is the person carrying out the test, as they are replicated from the real 94 samples.

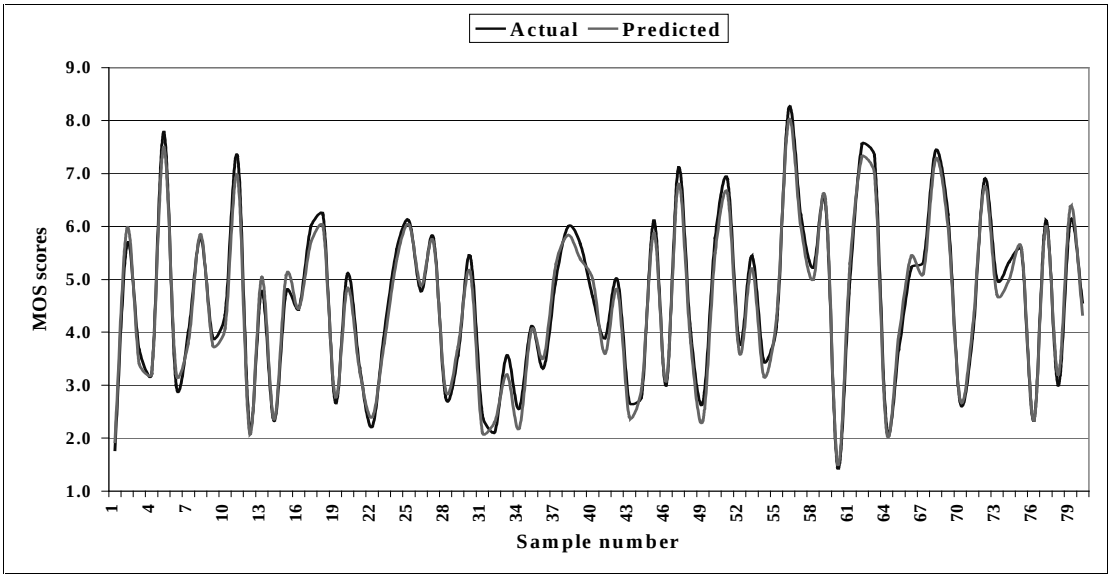

Figure 4. Actual and Predicted MOS scores for the training database

We invited 20 persons to perform the subjective tests. After that, a prescreening of the results was performed. As a result, we discarded the notes of two subjects, following the instructions in [15].

4) Training and testing the ANN

The number of input neurons in the input layer of our ANN is equal to the number of selected parameters (five). There is only one output neuron, the MOS score. The number of hidden neurons in the hidden layer is variable as it depends on the complexity of the problem (inputs, outputs, training set, and the global precision needed). In our case, with eight hidden neurons, we obtained the best results for both the training and the testing database together.

After carrying out the MOS experiment for the 94 samples, we divided our database into two parts: one to train the ANN containing 80 samples, and the other to test the ANN's accuracy to work in a dynamic environment, containing 14 samples. After training the ANN and comparing the training data against the values predicted by the ANN, we got a correlation factor = 0.9915, and average absolute error = 0.2084, showing that our model fits quite well the way in which humans rated the video quality. The result is shown in Figure 4.

5) How well does the ANN perform?

In order to address the question "How well does the ANN perform?", the ANN was applied to the testing set. The results were correlation coefficient = 0.9907 and average error = 0.253. Once again, the performance of the ANN was excellent, as can be observed in Figure 5.

From Figure 4 and Figure 5, it can be observed that the video quality scores generated by the ANN fits quite nicely the nonlinear model built by the subjects participating in the MOS experiment. Also, Figure 5 says that learning algorithms give neural networks the advantage of high adaptability, which allows them to self optimize their performance when functioning under a dynamical environment (that is, reacting to inputs never seen during the training phase).

6. Conclusion and future directions

In this paper, it has been described how ANN can be used to create a nonlinear mapping between non subjective audio signals measures (i.e., packet loss rate, loss distribution, bit rate, frame rate, encoded frame type, etc.), and a subjective (i.e., MOS) measure of video quality. This mapping mimics the way in which human subjects perceive video quality at a destination point in a communication network.

We have validated our approach by building the ANN model to assess in real time the video quality transmitted over the Internet, taking into account the previously mentioned parameters. We have shown that the ANN performs quite well in measuring video quality in real time.

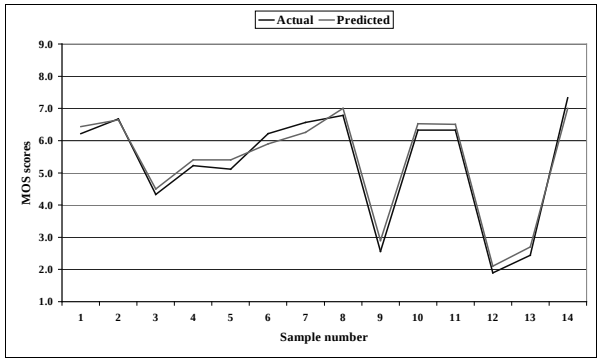

Figure 5. Actual and Predicted MOS scores for the testing database

As the video quality is affected by many parameters, one of the future directions in our research is to build a robust database, by conducting a series of MOS experiments taking into account different combinations of these parameters. The ANN approach allows also to identify the importance of network parameters in distorting video signals. Thus, using a tool that effectively measures video quality and identifies the nature of current distortions, better solutions to other problems would be developed, e.g., adaptive error correction schemes to dynamically compensate video distortion based on the current network situation, identification of the best trade-off between redundant information and bandwidth requirements to improve QoS, etc.

References

[1] Boyce, J.M., Gaglianello, R.D. "Packet Loss Effects on MPEG Video Sent Over the Public Internet", *Proc. ACM Multimedia'98*, 1998.

[2] Claypool, M, Tanner, J. "The Effects of Jitter on the Perceptual Quality of Video", *ACM Multimedia'99*, 1999.

[3] Cramer, C., Gelenbe, E., Gelenbe, P. "Image and video compression", in *IEEE Potentials*, 1998.

[4] Gelenbe, E., Bakircioglu, H, and Kocak, T. "Image Processing with the Random Neural Network", in *Proc. of the IEEE Digital Signal Processing Conf.*, June 1997.

[5] Ghinea, G., Thomas, J.P. "QoS Impact on User Perception and Understanding of Multimedia Video Clips", *Proc. ACM Multimedia 98*, 1998.

[6] Hands, D.; Wilkins, M. "A study of the impact of network loss and burst size on video streaming quality and acceptability" *Interactive Distributed Multimedia Systems and Telecommunication Services Workshop*, Germany, 1999.

[7] ITU-T Recommendation H.263 - Video coding for low bit rate communication, Feb. 1998. http://www.itu.int/

[8] Jones, C., Atkinson, D.J. "Development of Opinion-Based Audiovisual Quality Models for Desktop Video-Teleconferencing", *Interl. Workshop on Quality of Service*, 1998.

[9] Lavington, S., Hagras, H., Dewhurst, N. "Using a MLP to Predict Packet Loss During Real-Time Video Transmission", *Univ. Of Essex, UK, Internal Report CSM329*, July 1999.

[10] Le Leannec, F., Guillemot, C. "Packet Loss Resilient H.263+ Compliant Video Coding", *Interl. Conf. on Image Processing*, Sept. 2000.

[11] Mohamed, S., Cervantes, F., and, Afifi, H. "Audio Quality Assessment in Packet Networks: an Inter-Subjective Neural Network Model", *Proc. of Interl. Conf. on Information Networking*, Japan, 2001.

[12] Mohamed, S., Cervantes, F., and, Afifi, H. "Integrating Networks Measurements and Speech Quality Subjective Scores for Control Purposes", *Proc. of the IEEE INFOCOM*, April, 2001, Alaska.

[13] Mohamed, S., Afifi, H., Rubino, G., and Cervantes, F. "Video Quality Assessment in Packet Networks: a Neural Network Model", Technical Report no. 1400, IRISA, May 2001.
ftp://ftp.irisa.fr/techreports/2001/PI-1400.ps.gz

[14] Pessoa, A., Falcao, A., Nishihara, R., Silva, A., Lotufo, R. "Video Quality Assessment Using Objective Parameters Based on Image Segmentation", in *SMPTE Journal*, Dec. 1999.

[15] Rec. ITU-R BT.500-10. "Methodology for the Subjective Assessment of the quality of Television Pictures", March 2000.

[16] Rec. ITU-T P.910. "Subjective Video Quality Assessment methods for Multimedia Applications", September 1999.

[17] RFC 2429 "RTP Payload Format for the 1998 Version of ITU-T Rec. H.263 Video (H.263+)", IETF, 1998.

[18] Rumelhart, D.E., Hinton, G.E., and Williams, R.J. "Learning internal representations by error propagation". *Parallel Distributed Processing*, vol. 1. Cambridge, Massachusetts, MIT Press, 1986.

[19] Tan, W., and Zakhor, A. "Multicast Transmission of Scalable Video using Receiver-driven Hierarchical FEC" *in Packet Video Workshop 99*, April 1999.

[20] Watson, A. and Sasse, M.A. "Evaluating Audio and Video Quality in Low-Cost Multimedia Conferencing Systems", *Interacting with Computers*, 8(3), 1996.

[21] Watson, A. and Sasse, M.A. "Measuring Perceived Quality of Speech and Video in Multimedia Conferencing Applications". *Proc. of ACM Multimedai'98*, 1998.

[22] Wu, H.R., Ferguson, T. and Qiu, B. "Digital Video Quality Evaluation Using Quantitative Quality Metrics", *Proc. of the 4th Interl. Conf. on Signal Processing*, 1998.

[23] Wu, H.R., Lambrecht, C., Yuen, M., and Qiu, B., "Quantitative quality and impairment metrics for digitally coded images and image sequences", in *Proc. of Australian Telecommunication Networks & Applications Conf.*, Dec. 1996.