LA1 TestBed: Evaluation Testbed to Assess the Impact of Network Impairments on Video Quality

Mu Mu, Andreas Mauthe, John Casson, Gareth Tyson
Lancaster University
Lancaster, United Kingdom
{m.mu, andreas, g.tyson}@comp.lancs.ac.uk,
jcasson@gmail.com

Francisco Garcia
Agilent Laboratories
Edinburgh, United Kingdom
frankie_garcia@agilent.com

Abstract— **Currently, a complete system for analyzing the effect of packet loss on a viewer's perception is not available. Given the popularity of digital video and the growing interest in live video streams where channel coding errors cannot be corrected, such a system would give great insight into the problem of video corruption through transmission errors and how they are perceived by the user. In this paper we introduce such a system, where digital video can be corrupted according to established loss patterns and the effect is measured automatically. The corrupted video is then used as input for user tests. Their results are analyzed and compared with the automatically generated. Within this paper we present the complete testing system that makes use of existing software as well as introducing new modules and extensions. With the current configuration the system can test packet loss in H.264 coded video streams and produce a statistical analysis detailing the results. The system is fully modular allowing for future developments such as other types of statistical analysis, different video measurements and new video codecs.**

Keywords- Testbed, Network Impariments, Video Quality

I. INTRODUCTION

With the increasing numbers of live streaming internet channels, assessing how transmission impairments (namely packet loss) impact on the perceived video quality is becoming more and more important. This is because emerging high quality multimedia applications such as live streaming on IPTV are vulnerable to network impairments during content distribution. However, so far it has not been clearly establish how a breach in Quality of Service (QoS) affects the user perceived video quality. Since video content is processed by the application and its quality depends on various factors, traditional network QoS based quality metrics such as packet loss rate (PLR) are not capable of reflecting the impact of transmission errors on the end user's perception. The concept of Quality of Experience (QoE) has been introduced to address the issues concerning the assessment of how well a video service meets the consumers' expectations [1]. As an alternative, a number of video quality assessment models have been proposed [1-4] to address the various challenges. These include how to model the human visual system, user perception and how to assess different kinds of impairments. To evaluate the proposed models usually different testing scenarios are defined and subjective user tests performed to verify the correlation between the output of the objective models and the user test results. However, a complete system for evaluating the impact of impairments (such as packet loss) and comparing the performance of objective video quality assessment models with the viewers' perception of is not available. Due to the lack of such a testbed it is difficult to systematically establish what correlations exists between packet loss, video impairment and user perception. Further, the performance of different assessment models cannot be compared to each other and the flexibility of the models under various use case scenarios cannot be verified.

In this paper, we introduce the LA1 testbed, in which network errors can be introduced systematically according to well specified loss patterns and to measure and assess their impact on video sequences. It also allows evaluating the performance of objective video assessment models that measure the effect of the network impairments on video content. The LA1 testbed is designed to provide a common testing platform over which objective and subjective evaluations can be performed with different testing scenarios (e.g. codec, transmission pattern) in an automatic manner. In this paper, we describe the design and prototypical implementation of the LA1 testbed. A use case is also demonstrated.

The rest of this paper is organized as follows. In Section II, background and the state-of-the art of current video quality assessment approaches are presented. Related work on testbeds is discussed in Section III. Section IV introduces the design goal and the system level design of LA1. The implementation of LA1 is presented in Section V followed by a use case study in Section VI. Section VII summarizes and concludes the paper, giving an outlook on future work.

II. BACKGROUND

Traditionally video quality has been assessed by human users (sometimes trained and sometimes untrained) according to different testing procedures. However, with the explosive increase of audiovisual content this is not a viable way to establish if content adheres to a well specified service level anymore. Thus, a number of objective video quality assessment models have been developed that try to capture, model and reproduce the subjective testing in an algorithmic

manner. This is termed objective video quality assessment. Various models have been defined, however, it is still necessary for them to be verified against real users. Hence, special test procedures are necessary to carry out this verification in a systematic manner.

A. Objective Video Quality Assessment Models

Different types and classes of video quality metrics have been proposed in recent years [1-4]. Video quality metrics are classified using two orthogonal classifications: the amount of reference information required to assess the quality and the measured features [5].

1) Classification Based on the Amount of the Required Reference Information

Many research activities have focused on different methods of assessing quality considering the additional information used in the assessment process. Usually three approaches (metric classes) are distinguished. The first one is called the "Full-Reference" (FR) approach, which assumes unlimited access to the original (reference) video. Quality assessment is performed by comparing the distorted video to the original. Since full access to the original video is needed, the area of possible applications is restricted to laboratory tests. This includes codec testing and comparison, encoder tuning and quality acceptance level testing. Examples of full-reference metrics are MPQM [6] and SSIM [7].

The second class is commonly referred to as "No-Reference" (NR) which involves quality evaluation without any knowledge of the original material. Quality assessment is performed in a no-comparative manner since the original video content is not available. It is based on assessing well specified features and parameters. The goal is to use these methods to build services enabling such things as real-time in-service assessment for network performance monitoring, alarm generation in case of severe quality deterioration and quality-based billing. The performance of NR models is limited by the knowledge available about the target video content. If only the basic network delivery information, such as packet loss rate is available, the NR model shows low correlation to the user perception. However, if slightly more information can be taken into account (such as codec or video type) the NR performance can be improved.

The last class is referred to as "Reduced-Reference" (RR) approach which merges the advantages of both FR and NR approaches. Only some well specified features (such as motion information or certain spatial details) are extracted from the reference video stream and used for comparison. This information is usually communicated out-of-band, and leads to a partial alignment of compared parts. The idea is that only key features are used so that the amount of extracted information is still manageable to allow more precise quality evaluation. RR is used in both laboratory and in-service scenarios. At present there are only few video quality metrics using this approach, namely [8] or [9].

2) Classification Based on Measured Features

Ultimately the quality of the *presented* video is the only relevant factor. If only this is assessed the whole end-to-end video delivery system is considered as a black box and only the decoded video quality at the receiver side (in a comparative or an absolute way) is analyzed. Thus, this process includes an assessment of the overall video quality which does not discriminate between the different kinds of impairments or where they might have been generated. This kind of assessment is commonly referred to as "artifact measurement" (AM). It does not take into account what might have happened at which stage of the video processing life-cycle, i.e. where quality loss might have occurred, what kind of loss has happened and if this has happened in a random or moderated fashion.

The second approach is much more focused on the transmission system and considers relevant parameters of the delivery system which are collected in order to predict video quality. The knowledge about the kind impairment and the way it can be introduced is used to optimize the assessment. This approach is referred to as "quality of delivery" (QoD). Research has been recently carried out in this context [10-12] that considers network QoS parameters as well as extended application/user level configurations (such as video characteristics and user preferences).

B. Subjective User Test

In order to verify the performance of the objective models, subjective user tests are usually performed to study the correlation between results achieved by the objective models and the user opinion. The methodology for the subjective assessment of the quality of television pictures is recommended in [13]. Apart from common features such as viewing conditions, test material selection and test instructions, three subjective test methods are further described in [13], namely Double-stimulus continuous quality-scale (DSCQS), Double-stimulus impairment scale (DSIS) and Simultaneous double stimulus for continuous evaluation (SDSCE). Absolute category scale (ACR) is a single-stimulus method that is proposed in [14]. The ACR test procedure includes only one reference version of each video sequence, not as part of a pair, but as a freestanding stimulus for rating like other corrupted ones [15]. In [16],some alternative methods have also been studied. It is recommended that in practice, particular methods should be adopted to specific problem spaces i.e. they should be used to address particular assessment problems [13]. These kinds of user tests have to be specially carried out to establish if a newly devised method performs according to specification. It is important that they are performed according to given standards to ensure that they are reproducible.

III. RELATED WORK

In the AQUAVIT project [17] testbeds have been specifically designed to allow assessment methods for audio-visual transmission over IP and UMTS networks to be studied.

The testbeds use the most relevant audio-visual compression algorithms and allow the simulation of transmission over UMTS and IP networks in the presence of bit errors and packet loss. Two testbeds were designed for different testing scenarios: (i) an IP testbed that can operate both in real-time and non-real-time and (ii) a UMTS testbed, including a UMTS-channel simulator, which operates in non-real-time only.

EvalVid is a complete framework and tool-set for the evaluation of the quality of video transmitted over a real or simulated communication network [18]. Network QoS parameter calculation as well as a video quality evaluation of the received video based on the PSNR calculation is also supported. Within EvalVid different networks and codec scenarios can be used. Video Sender (VS), Fix Video (FV) and Evaluate Traces (ET) are the three main components of the EvalVid framework. The Video Sender (VS) generates a trace file from the encoded video file. The results produced by VS are two trace files containing information about every frame in the video file and every packet generated for transmission [18]. These two trace files together represent a complete video transmission (at the sender side) and contain all information needed for further evaluation by EvalVid. The actual calculation of packet loss, frame loss and delay/jitter is performed by Evaluate Traces (ET). For the calculation of this data only the three trace files (sender trace, receiver trace and video trace) are required. Another task ET performs is the generation of a corrupted video file (due to losses). This corrupted file is needed later to perform the end-to-end video quality assessment. Thus, the original encoded video file is needed as an input for ET. The Fix Video (FV) component simply reorders, decodes and reconstructs the YUV raw pictures, i.e. it is not an actual evaluation component.

Although the current testbed designs are mostly open to be used with various application and network testing scenarios, the support of objective models are limited due to their architectural design. For example, some of the network based (i.e. QoD) quality assessment models rely on the video metadata information to a relatively large extent. Thus, the evaluation of these models requires specific data support from the underlying testbed.

IV. System Design

A. Design goals

Before developing the LA1 testbed we have extensively studied the requirements in order to establish the basic features and principles a testbed for the assessment of the impact of network impairment on video quality should have. This has resulted in the following set of design goals such a system has to fulfill:

- Support of different assessment models i.e. NR, RR and FR
- Support both network based and video measurement based assessment models
- Flexible definition of various application scenarios taking into account different use cases
- Support for network simulations to model different network types and loss patterns
- Integration of basic data/statistic analysis to compare the results of different approaches

These design goals have subsequently guided the actual system design process.

B. System architecture and design

The architecture of the LA1 test bed is shown in Figure 1. The testbed takes a *test video sequence* and *test configuration parameters* as input and gives data analysis results as output. The model comprises seven function blocks, namely *Encoder*, *Packet Loss Simulator (PLS)*, *Decoder*, *Video Analyser*, *Objective Quality Assessment Models (OQAM)*, *Subjective Experiments* and *Statistic Analysis*. In the following, input, output and function blocks of the testbed are described in more detail.

1) Input video sequences

Uncompressed YUV video sequences are the first input component to the testbed. The YUV format and frame size can vary for different test scenarios but must be indicated in the test configuration file. If it is required for professional blind benchmarking the target video sequences must be kept in a secret database. The test video can then be selected randomly from the database during the test procedure. Certain metadata such as genre (e.g. sports or cartoon) may also be available with the video sequences.

2) Output

The output of the testbed after the entire test procedure has been concluded is the statistical analysis of the results. This includes not only the objective quality assessment results but also associated data from subjective experiments. The definition and scale of the output are dependent on the statistic models that are adopted for the test. Apart from the final output, all the raw test data and intermediate log files will be kept for potential extended studies. For example, psychology study can be performed on the subjective user test results.

3) Configuration file

The configuration file contains parameters that define the test scenarios and conditions. Three sets of testing parameters are included in the configuration file: *encoding configuration*, *packet loss configuration* and *decoder configuration*. In the encoding configuration, video files are set to a predefined use case scenario (e.g. Mobile IPTV) which determines factors such as screen size, video framerate, bitrate, error control etc. The decoder configuration gives the capacity (e.g. buffer size) and capability (e.g. error concealment mechanisms) of the pre-defined end device. The packet loss configuration sets up the specified loss pattern so the packet loss simulator can simulate the corresponding networking transmission conditions.

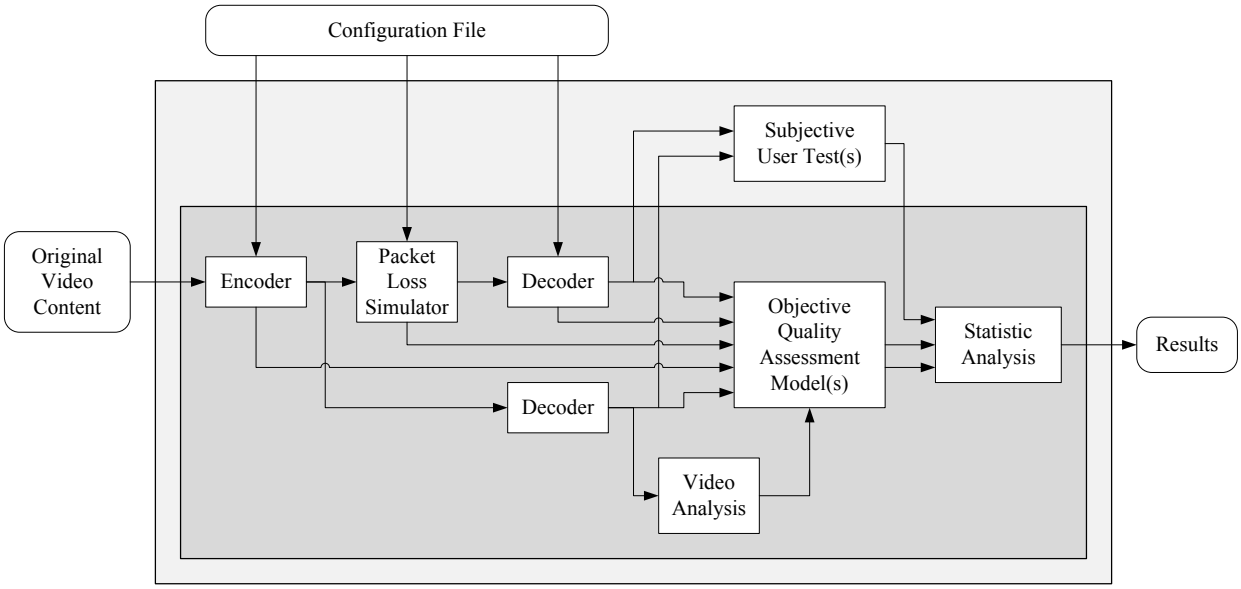


Figure 1 LA1 Testbed: Evaluation Testbed to Assess the Impact of Network Impairments on Video Quality

4) Encoder

The encoder encodes the target video sequences considering encoding parameters such as framesize and framerate from the encoding configuration file. The outputs of the encoder are the reference video sequences and the encoding log file. The encoding log file contains information on how exactly each frame is coded, compressed and packetized. The network simulator can use the log file to perform the packet loss on frame types or specific spatial location of the frames. Some of the models in OQAM also require the encoding log file to estimate the perceptual deterioration from the spatial and temporal distribution of packet loss.

5) Packet Loss Simulator (PLS)

The packet loss simulator removes video content from the encoded video sequence (on a network packet basis) to simulate the packet loss in the delivery network. The removal scheme is indicated in the configuration file and can be set to three modes: (i) random, (ii) pattern specific, and (iii) packet specific. When the removal scheme is set to random, the PLS discards content within a well specified period (also indicated in the configuration file) using a loss probability model such as Markov chains. In pattern specific mode, packet loss is simulated according to testing scenarios (e.g. P-frame only or loss pattern in wireless mesh networks). The PLS may require the encoder log file to associate packets with specific video frames. In the packet specific mode, the PLS looks into the sequence number of the packets to be simulated as lost packets (the sequence number can be used directly or after random shifting to generalize the simulation). The corrupted video content from PLS is then sent to the decoder. The loss simulator log file is also generated and made available to the OQAM. This is since some of the assessment models under test are based on the network monitoring results.

6) Decoder

The decoder(s) decodes both the reference video sequences and corrupted video sequences back to the original video format (i.e. the input format). When decoding corrupted video sequences the decoder simulates the user terminal's decoding progress according to the decoder parameters in the configuration file. More than one instance of the decoder can be used in the testbed (taking into account run-time performance restrictions). However, the decoder parameters must be the same across all the decoder instances. This is in order to avoid decoder bias on different video sequences. A decoder log file is also created after the decoding progress has been concluded. This log file contains the actions the decoder does which may impact on the packet loss effects of the decoded video.

7) Video Analysis

Some quality assessment models consider application/user level information (such as video content characteristics) as one of the main factors that affect the perceived packet loss impact on video content. Thus, it is important to consider them within the testbed. In a real-world scenario these content characteristics (such as motion and complexity level of the video frames) would be provided by the actual content provider or an abstraction function of the encoder in a well specified manner. In our testbed the video analysis model provides the video characteristic information by image/video signal processing on the decoded video sequences. The analysis is carried out on the reference video to avoid impact from the packet loss that is simulated in the PLS. Alternatively, this function may also be supported within the encoder during the encoding process.

8) Objective Quality Assessment Models (OQAM)

Depending on the type of assessment model employed different sets of inputs to the OQAM are enabled. For full-reference artifact measurement models, the decoded reference video sequences and decoded corrupted video sequences are both available. For no-reference artifact measurement models, only decoded corrupted video sequences will be provided as input. For the network measurement based models the loss simulator log is fed into the system alongside the encoder/decoder log file. Further, the video analysis results can be also provided. The OQAM generates user experience estimation as output.

9) Subjective user test

The subjective user test takes decoded corrupted video sequences and/or decoded reference video sequences as testing material. Features and methodology of the user test must be chosen to address particular assessment scenarios. Subjective user testing is not part of the autonomous system in the testbed (Figure 1), thus the user test configuration must be recorded manually for further studies. The output of subjective user tests is described numerically in qualitative scales (e.g. MOS or Differential MOS). The final score of a video sequence represents the average value of the scores provided by all participants viewing the video sequence. If a continuous scoring method such as SDSCE [13] is performed, an overall score of a test sequence must be derived from each participants before summarized with scores from other participants.

10) Statistic Analysis

Experiment results from the OQAM and subjective user tests are analyzed with statistic methods. All the inputs to statistic analysis must be in the same scale so data pre-processing is required confronting the mismatch between OQAM results and user test results.

As was summarized in [19], the performance of each objective quality model is characterized by three prediction attributes: accuracy, monotonicity and consistency. These three attributes are quantified by the root mean square (RMS) error, Pearson correlation and the outlier ratio. The objective quality model evaluation was performed in three steps. The first step is a mapping of the objective data to the subjective scale. The second calculates the evaluation metrics for the models and their confidence intervals. The third test is used to establish statistical differences between the evaluation metrics value of different models [19].

V. IMPLEMENTATION

This section introduces a prototypical implementation of the LA1 testbed to benchmark objective quality assessment models. It focuses on a system using H.264/AVC video sequences that are corrupted by well specified packet loss in delivery.

From Figure 1 it can be seen that Encoder, Packet Loss Simulator and Decoder are the main test preparation functions in the LA1 testbed. The H.264/AVC JM Reference Software [20] is currently the only full implantation of H.264 codec used. An RTP packet loss simulator is also provided as part of the JM Reference software. However, extended functions such as calibration between lost packets and frames, loss pattern integration and logging are not available from the original software. In the following sections, the necessary extensions to the reference software that have been made to achieve the design goals of LA1 that are described. The implementation of the Video Analysis, Subjective Experiments and Statistic Analysis function block will also be presented.

A. Encoder/Decoder

The modifications to the encoder/decoder source code enable the program to output information about the occurred data corruption to a log file. The first enhancement in this context is to open a corruption log file in an appropriate section of the program such that any other function may write to this file. This is achieved by defining a global variable in the main source code file, then opening the file for writing. This global variable can then be used to write data to the log file throughout the program.

a) corruptionLog.txt

This file outputs results by the decoder and provides detailed information about the last video clip decoded. For each frame decoded, the following information must be present in the file:

- A line showing that a new frame is being decoded:

  "Entering frame number %i"

  Where %i is the integer number of the frame currently being decoded

- A line showing that decoding of a frame has completed:

  *"*** End of Frame %i Type:%s"*

  Where %i is the number of the frame that has finished decoding and %s denotes the type of frame.

If a corruption has occurred within the frame:

- A line detailing the sequence number of the lost RTP packet in the following format:

  "RTP Packet %i lost"

  Where %i is the sequence number of the lost RTP packet

- A line in the following format, detailing each lost macroblock:

  "One MB Lost in frame %i: co-ords: %i %i"

  Where the first %i is the frame number with the lost macroblock, the second %i is the x co-ordinate of the top corner of the macroblock lost and the third %i is the corresponding y co-ordinate.

b) Breakdown.txt

This file contains the results recorded by the decoder on successful completion of decoding a video file. The purpose of this file is to show the makeup of a H.264 encoded stream e.g. which macroblocks are in which frames and which RTP packets make up these frames. This is used to debug the system and to provide information to assist with the decision as to which packets to drop.

The file is in the following format:

- When a new RTP packet is passed to the decoder, a line is written to the file in the following format:

  " New RTP packet, number %i"

  Where %i is the sequence number of the RTP packet

- When starting to decode a new frame, a line in the following format is written to the file:

  *"*** POC %i Type: %s ***"*

  Where %i is the current frame number and %s is the type of frame

- When a new macroblock is being decoded at the start of the frame the following line is added:

  "MBs: %i %i %i"

  Where the number of %i's in the line is variable and each corresponds to the macroblock currently being decoded.

- When a new macroblock is being decoded not at the start of the frame, with an indent of five characters in the following format:

  " %i %i %i"

  Where the number of %i's in the line is variable and each corresponds to the macroblock currently being decoded.

A breakdown logfile is shown in Figure 4. With the breakdown file, the calibration between frame number, frame type, packets and macroblocks established. The calibration is critical for the testbed in order to evaluate the performance of some network based video assessment models, which require application metadata for the analysis. The calibration map of the B frame in Figure 5 is showed in Figure 4.

B. Packet Loss Simulator

In its original form, the RTP loss simulator drops packets from a file using two parameters: an initial offset and a percentage of packets lost. The percentage packet loss drops random packets using the percentage value as a seed resulting in the same packets being dropped if the same percentage value is used. The modifications to this program will improve on this loss model by adding the following features:

- Ability to drop specified packets using a list
- A seeded random number generator offset for packets to be dropped

This allows a list of packets to be given and different packets to be dropped in each run of the simulator.

C. Data Analysis

Various statistical analysis models can be adopted as part of the Data Analysis function block. For example, to calculate the correlation between the results of two models the following implementation is processed.

The Pearson product moment correlation coefficient gives a value that shows the correlation between two random variables. If the variables are "x" and "y", and "n" is the number of pairs of data then the correlation coefficient can be calculated using the following formula:

$$r = \frac{n\sum xy - \left(\sum x\right)\left(\sum y\right)}{\sqrt{n\left(\sum x^2\right) - \left(\sum x\right)^2}\sqrt{n\left(\sum y^2\right) - \left(\sum y\right)^2}} \quad (1)$$

"r" is a value between 1 and -1 where the magnitude of the value shows either positive or negative correlation. If "r" is close to 0 then the values show little to no correlation. As "n" increases, the significance of the result increases. For example, a correlation coefficient of 0.8 is not a significant result if "n" is 10, however if "n" is 50 then a coefficient of 0.85 carries a much greater significance. This is an important point to note as it indicates that a greater number of tests will produce a more significant outcome.

Our tool which implements the Pearson correlation opens a list where two values are stored per line and calculates the correlation coefficient for these sets of values. The final correlation coefficient is then output.

VI. Experimental Evaluation

In this section, an initial experiment with the LA1 testbed is described to demonstrate its operation. The scenarios used for these experiments follow the ones used in the design phase. In the experiment one video sequence is fed into the testbed. The encoding, decoding as well as a loss configuration file (which indicates the loss pattern to be simulated) are defined in the testbed configuration file. Both the decoded reference video sequence and corrupted video sequence are sent to the objective video quality assessment model function block where two models evaluate the corrupted video sequence. The data analysis function block then calculates the correlation between the results of two assessment models.

A. Configuration file

Figure 2 shows the example of a test configuration file which comprises encoder configuration, decoder configuration and packet loss simulator configuration. Framesize, framerate, frametype, compression level, error control mode and other encoding parameters are defined in the encoder configuration. In the decoder configuration of this experiment, only the decoder error concealment parameter is defined (all the other

decoder parameters are kept as specified by their default values). The packet loss simulator configuration defines the loss pattern to be simulated on the target video sequences. In this experiment shown here the loss of three packets (number: 6, 30 and 55) has been simulated.

```
**TESTBED CONFIGURATION:
*ENCODER CONFIGURATION:
InputFile              = "walk_qcif.yuv"
StartFrame             = 0
FramesToBeEncoded      = 5
FrameRate              = 30.0
SourceWidth            = 176
SourceHeight           = 144
SourceResize           = 0
LevelIDC               = 40
IDRPeriod              = 3
QPISlice               = 28
QPPSlice               = 28
FrameSkip              = 1
NumberReferenceFrames  = 5
PSliceSkip             = 1
NumberBFrames          = 1
QPBSlice               = 30
QPSPSlice              = 36
SliceMode              =  2
SliceArgument          = 300

*DECODER CONFIGURATION:
Err Concealment        = 0

*PACKET LOSS SIMULATOR CONFIGURATION:
6
30
55
```

Figure 2 Test configuration

B. *Models under test*

The tests of the prototypical LA1 testbed looks specifically at two different objective video assessments, i.e. Weighted Peak Signal to Noise Ratio (WPSNR) and the "Number of Lost Pixel Model". At this stage no objective video quality assessment has been considered. However, the principles of how to integrate them and how to use their results remain the same.

1) *Weighted Peak Signal to Noise Ratio - WPSNR*

The peak signal to noise ratio is often chosen as the primary measurement value. Despite the fact that it often does not very closely reflect the users' quality assessment of a distorted video, it is a commonly used measure. In order to improve its performance we have implemented a version that takes the position of a visual distortion within a frame into account since this should better reflect how an end user perceives images distortions and therefore rates image quality. The hypothesis applied in this context states that if the image is separated into three equally sized horizontal rows, distortions in the centre row would be perceived as being worse than the same distortions in either the top or the bottom row. Through this division the region of interest is modelled, i.e. the region within an image a user pays closer attention to. While the division used for WPSNR does not exactly reflect the region of interest (since this also heavily depends on the actual video content), it does provide a better approximation than the non weighted PSNR model.

The WPSNR value is calculated by splitting the image into nine equally sized areas, computing the PSNR values for each area then averaging these PSNR values. However, rather than calculating the mean average, the three sections making up the centre horizontal row are weighted in the calculations, giving them greater influence over the final value. The centre values are weighted by multiplying them by a value greater than one before they are summed up. Subsequently the final result is divided by the result of the sum of the weights. The following figure (Figure 3) shows how the image is split and the weights for each section:

Figure 3 Weight coefficients of WPSNR

The PSNR value for each section is multiplied by the respective section weight. This value is then added to each of the other WPSNR values. The final WPSNR calculation is made by dividing the sum total of the weighted values by the sum of the weights. In this case, the sum is thirteen. This can be adjusted and the division of the image into different section is configurable. Thus, the WPSNR model can be improved by using more sections and discriminating further between them. More tests have to be carried out to establish what the most optimal division is. This will also take into account different content types to achieve an as accurate approximation as possible.

2) *Number of Lost Pixels*

The "Number of Lost Pixels" model simply measures the number of pixels that have been lost due to the packet loss. This model is supported by the packet-macroblock calibration map as shown in Figure 5. Although it quickly indicates the coverage of the packet loss in video frames, the "number of lost pixels" model neglects the effect of error concealment with which the lost pixels can be estimated by received content.

C. Results

A breakdown logfile is shown in Figure 4. With the breakdown file, the calibration between frame number, frame type, packets and macroblocks is established. The calibration is critical for the testbed to evaluate the performance of some network based video assessment models which requires application metadata for the analysis. The calibration map of the B frame in Figure 5 is showed in Figure 4.

```
"breakdownLog": (excerpt)

*** POC 6 Type:B ***
MBs: 0 1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16
      17 18 19 20 21 22 23 24 25 26 27 28 29 30
   New RTP packet, number 60
     31 32 33 34 35 36 37 38 39 40
   New RTP packet, number 61
     41 42 43 44 45 46 47
   New RTP packet, number 62
     48 49 50 51 52 53 54 55
   New RTP packet, number 63
     56 57 58 59 60 61 62 63 64
   New RTP packet, number 64
     65 66 67 68 69 70 71 72 73 74 75 76 77 78
   New RTP packet, number 65
     79 80 81 82 83 84 85 86 87 88 89 90 91 92
      93 94 95 96 97 98
   New RTP packet, number 66

*** POC 0 Type:IDR ***

MBs: 0 1 2 3 4 5 6 7 8 9 10
   New RTP packet, number 67
     11 12 13 14 15 16 17
   New RTP packet, number 68
     18 19 20 21
   New RTP packet, number 69
     22 23 24 25
   New RTP packet, number 70
     26 27 28
   New RTP packet, number 71
      ...
```

Figure 4 Breakdown log

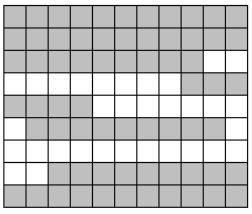
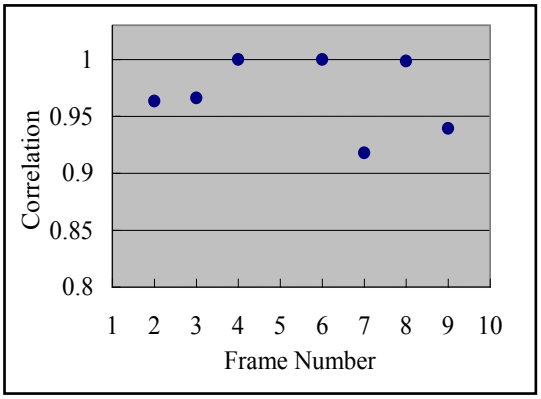


Figure 5 Packet loss-macroblock loss calibration

Figure 6 depicts the output of the testbed which shows the correlation between the WPSNR values and the number of lost pixels in each frame. The video frames number 1, 5 and 10 were not corrupted (so the values of these frames are shown as "NaN" in Figure 5 and eliminated in Figure 6) as a result of the randomness of dropping packets. It can be concluded that the outputs from the two models reach the highest correlation on frame number 6 and the lowest correlation on frame number 7. The difference between the correlation on each video frames reflects the different spatial areas that the packet loss affects across different frames. On frame number 7, the deteriorations concentrate mostly on the centre of the video frame where the weight coefficient is the higher than the rest of the frame.

Frame#	Correlation
1	NaN
2	0.9634159446540651
3	0.9660568635548616
4	0.9998551995392464
5	NaN
6	0.9999754925913164
7	0.9179433996037352
8	0.9985659314304537
9	0.9393823372599739
10	NaN

(a)



(b)

Figure 6 Correlation between two models under test

Further tests are necessary to establish how closely these results correlate with the user perceived video quality. However, this evaluation shows that we have achieve our goal of creating and open and flexible testbed through which we can test different objective video quality assessment models. Even more important in this context is that it can also be used to further study the impact of network impairments on the user perceived quality in much more detail. LA1 is an efficient tool that will help us to systematically study this topic and also offers an efficient and open way to establish the different influencing factors in this context. Thus, through LA1 we aim to determine if a systematic correlation exists and what the determining causes are.

VII. Conclusion and Future Work

In this paper the LA1 testbed is introduced. LA1 is based on a modular framework that allows assessing the impact of network impairment on videos and studying different assessment models as well as conducting user tests. More specifically, the LA1 testbed is designed to simulate packet loss that reflects real network scenarios. This testbed enables us on the one hand to evaluate objective video quality assessment models and their performance of measuring quality of video sequences facing packet loss within the content

delivery path. On the other hand LA 1 also allows us to conduct user tests that help to establish a correlation between network impairments and perceived video quality. The user tests can also be used to test the effectiveness of the objective video quality assessment models. Thus, the LA1 Testbed is designed to provide a common testing platform on which objective and subjective evaluations can be performed with different testing scenarios (such as codec, transmission pattern) with the least human intervention (in terms of set-up and analysis).

An initial implementation of the LA1 testbed has been realized and experiments with two objective video quality assessment models are used to demonstrate its effectiveness. The LA1 testbed framework has also been employed in our recent work [12, 21] to investigate the relationship between user perceived video quality and network impairments and hence facilitates the design of assessment models capturing all relevant factors in this context.

As a next step more function blocks representing different objective video assessment models will be implemented. The goal is to create an open platform for the comparison and evaluation of such models. Further, LA1 will be used to conduct user tests on a large scale. Ideally these would be replicated throughout partner institutions so that we can gather the rich data set necessary to conclusively establish the link between network impairments and user perceived video quality. All these tests will be carried out according to well specified test procedures [12, 13].

VIII. ACKNOWLEDGEMENT

The work presented in this paper is supported by the European Commission, under the Grant No.FP6-0384239 (Network of Excellence CONTENT) and Agilent Laboratories UK.

REFERENCE

[1] S. Winkler, "Video quality and beyond," in *European Signal Processing Conference*, Poland, 2007.

[2] A. M. Eskicioglu and P. S. Fisher, "Image quality measures and their performance," *IEEE Transactions on Communications,* vol. 43, pp. 2959-2965, December 1995.

[3] Y. Wang, "Survey of objective video quality measurements."

[4] Z. Wang, "Objective image/video quality measurement – a literature survey," *EE 381K: Multidimensional Digital Signal Processing*.

[5] P. Romaniak, M. Mu, A. Mauthe, S. D'Antonio, and M. Leszczuk, "Framework for the Integrated Video Quality Assessment," in *18th ITC Specialist Seminar on Quality of Experience*, Blekinge Institute of Technology, Karlskrona, Sweden, 2008.

[6] C. J. Van den Branden Lambrecht and O. Verscheure, "Perceptual quality measure using a spatiotemporal model of the human visual system," *Digital Video Compression: Algorithms and Technologies,* 1996.

[7] Z. Wang, L. Lu, and A. C. Bovik, "Video quality assessment based on structural distortion measurement " *Signal Processing: Image Communication,* vol. 19, 2004.

[8] S. Kanumuri, P. C. Cosman, and A. R. Reibman, "A Generalized Linear Model for MPEG-2 Packet Loss Visibility," in *Packet Video Workshop*, December 2004.

[9] S. Wolf and M. H. Pinson, "Spatial-temporal distortion metrics for inservice quality monitoring of any digital video system," *Proc. SPIE,* vol. 3845, pp. 266-277, 1999.

[10] J. Kangasharju, M. Mu, and G. D. Colussi, "Application-Level Fairness," in *International Conference on Information Networking 2008 (ICOIN 2008)*, Busan, Korea, JAN 2008.

[11] M. Mu, A. Mauthe, and F. Garcia, "A Utility-based QoS Model for Emerging Multimedia Applications," in *First IEEE Future Multimedia Networking (FMN 08) Workshop*, Cardiff, UK, 2008.

[12] M. Mu, R. Gostner, A. Mauthe, F. Garcia, and G. Tyson, "Visibility of Individual Packet Loss on H.264 Encoded Video Stream – A User Study on the Impact of Packet Loss on Perceived Video Quality," in *Sixteenth Annual Multimedia Computing and Networking (MMCN'09)*, San Jose, California, USA, 2009.

[13] "Methodology for the subjective assessment of the quality of television pictures," *ITU Recommendation BT.500-11*.

[14] "Subjective video quality assessment methods for multimedia applications," *ITU-T Recommendation P.910*.

[15] "Hybrid Perceptual/Bitstream Group TEST PLAN," *Video Quality Expert Group,* 2008.

[16] "Studies toward the unification of picture assessment methodology " *ITU Recommendation BT.1082*.

[17] "Project P905-PF EURESCOM. AQUAVIT - Assessment of QUality for Audio-Visual signals over Internet and UMTS," 2000.

[18] J. Klaue, B. Rathke, and A. Wolisz, "EvalVid - A Framework for Video Transmission and Quality Evaluation," in *13th International Conference on Modelling Techniques and Tools for Computer Performance Evaluation*, Urbana, Illinois, USA, September 2003, pp. 255-272.

[19] "Final report from the Video Quality Expert Group on the validation of objective models of multimedia quliaty assessment, Phase 1," *Video Quality Expert Group,* 2008.

[20] "H.264/AVC JM Reference Software," http://iphome.hhi.de/suehring/tml/.

[21] M. Mu, A. Mauthe, G. Tyson, E. Cerqueira, and F. Garcia, "The Impact of Network Impairments on the Perceptual Quality of IPTV Service," *In submitting,* 2008.