CONTRIBUTION TO IEEE STANDARDS SUBCOMMITTEE

Document Number:

DATE: August 5, 1997

COMMITTEE: G-2.1.6 Compression and Processing Subcommittee

TITLE: Discussion Document for the ITU-T JRG Topics

EDITOR: David Fibush

SOURCE: Tektronix Inc.

CONTACTS: David Fibush
Tektronix Inc., MS 50-353
P.O. Box 500
Beaverton, OR 97077
(503)627-6289
(503)627-1707 (fax)
davef@tv.tv.tek.com

DISTRIBUTION: G-2.1.6 Compression and Processing Subcommittee

ABSTRACT: This document lists discussion topics suggested by Arthur Webster for the ITU-T JRG and the proposed responses by Tektronix. They are offered for discussion at the IEEE G-2.1.6 subcommittee meeting to be held August 4th.

NOTICE

Discussion Document

IEEE and ANSI Meetings, August 4 and 5, 1997

ITU-T Joint Rapporteurs Group Topics

This document lists discussion topics suggested by Arthur Webster for the ITU-T JRG and the proposed responses by Tektronix. They are offered for discussion at the IEEE and ANSI T1A1.5 meetings to be held August 4th and 5th, respectively. Topics are shown in san-serif type and responses are shown in serif type.

2. Test plan(s) for Subjective and Objective Test(s). These tests should be video only. Some items to be addressed for the MPEG2 test:

a. List of HRC's (Hypothetical Reference Circuits) (which means end-to-end systems e.g. MPEG2 @ 2.0Mbps with bit errors)

2a. Hypothetical reference circuits (HRCs) will need to be related to the classes of operation (see item 4 for a definition of the classes). It is expected some HRCs may cover all classes while others will cover a range of classes or perhaps only one class.

i. How many HRC's? (10?)

2a-i. We would certainly like to limit the number of HRCs. Ten may be too small, perhaps 15 as a maximum to limit the size of the effort. Variables in the choice of HRCs may be: bitrate, generations of compression, additive/proportional noise, analog/digital concatenations.

ii. Limited to MPEG2?

2a-ii. MPEG-2 is not appropriate for classes 4, 5 and 6, H.261 and H.263 are required. See items 3 and 4.

b. List of Test Scenes

2b. We need to first determine the attributes of the test scenes. Again, some are class dependent similar to HRCs.

i. How many scenes? (10?)

2b-i. Ten will not be enough scenes considering the broad range of classes, perhaps as many as 15-20 will be needed.

ii. Length (10sec?)

2b-ii. Ten seconds is a good length to provide for subjective evaluation. We need agreement on consistency of material throughout the scene. It would be best to apply the objective methods to the complete set of test scenes without any training, but if some methods require training sets, than the training set should be limited to less than 10-20% of the test set. This ensures that we measure the generalizability of the objective method and not its ability to fit to the specific data.

iii. Format (e.g. 601)

2b-iii. The I/O should be restricted to Rec 601. A significant problem is line standard. Doing all scenes in both 525 and 625 will double our work. Our suggestion is to do some tests in each standard much as was done with the 4:2:2 profile evaluation, with the goal of balanced equal coverage of 525 and 625 material.

c. List of Labs willing to perform subjective and/or objective testing.

2c. Tektronix can help with objective testing, and possibly also with video preprocessing of the scenes (e.g. removing picture shifts, etc) based on available time and resources.

d. Analysis methods. (Mr. Beerends offers 2 options)

2d. The choice of model should be based on a number of items important to robust operation of an objective quality method. The method should ideally provide monotonic agreement with picture degradation, be generalizable over a broad set of scenes, and avoid serious breakdowns in performance for peculiar scenes or types of scenes. Robust behavior should be the most important characteristic as we cannot predict with any certainty the actual types of video that that a user may analyze with the method. A mathematical metric should be selected that suitably takes into account: correlation, monotonicity, and consistently reliable prediction. In addition to an agreed mathematical metric for objective model performance, we need to look at the results in the same manner as a customer who has purchased a measurement instrument. Do the data points make sense? In particular we need to look at data points where the objective and subjective measurements disagree and understand whether the failures are explainable.

e. Production of video data. Source scenes and compressed scenes.

2e. It would be best to obtain scenes that already exist and apply degradations per our selected HRCs. All input and output scenes will be Rec 601. Use of compressed scenes as source material is discouraged because of differences in decoders.

f. Division of data into training and testing sets.

2f. It is important to have testing set cover the full range of applications.

g. Subjective test method?

2g. The subjective test method should be DSCQS with an expanded scoring scale. Double stimulus because that corresponds directly to the objective measurement of picture degradations. Expanded scale to reduce the possible non-linear scoring compression at the high quality end.

h. Other issues (related to the comments of Mr. Fibush and Mr. Schertz)

i. Do we agree to limit the test to all digital paths? Or do we allow analog portions? If we allow analog portions then how do we ensure exact spatial alignment?

2h-i. Analog portions of the HRC should be allowed as these are relevant to real-life situations. Post-processing of the HRC results should be used to provide realignment. Composite (NTSC/PAL) should be considered part of the HRCs, but probably only for less than 10-12 Mbit/sec compressed scenes because it masks compression artifacts at MPEG-2 bit rates above about 12 Mb/s.

ii. If we limit a test to MPEG2 then frame repetition (i.e. dropped frames) will not be a major concern-- but it may be present with some HRC's. If frame repetition occurs, then it is possible that variable video delays will be introduced. If so, how will this be handled.

2h-ii. We have agreed to consider systems that do, and systems that do not, drop/repeat frames. In systems that do not drop/repeat frames (broadcast) such results are considered a broken system and precise objective measurements are not needed. In the other systems (video conferencing) the statistics of the dropped/repeated frames and timing variations should be measured along with the resulting picture quality (degradation).

iii. If the selected HRC's exhibit changes in chroma and luma gain, black level, chroma/luma differential timing, possibly based on scene content, then how will this be handled?

2h-iii. Changes in chroma and luma gain, black level, and chroma/luma differential timing may come from two sources; system misadjustment which is the same for all pictures or changes which are a function of picture content. The former are systematic errors which should be removed by post-processing just as done with spatial alignment (this is exactly what will be done in the display for the DSCQS measurements). The latter are compression artifacts which should be included in the subjective and objective measurement results.

iv. If Sync/burst jitter (picture jitter) is present how should it be handled? (If visible to the viewer

it should be left in, if not visible, then it could be calibrated out.)

2h-iv. Any jitter which is visible is due to bad design or incorrect system setup. It should not be part of, or considered in, these tests.

v. What transmission errors should be included?

2h-v. It has been suggested that some HRCs should include application appropriate error rates, and we agree with the relevance of this item, but a useful selection of error rates is likely to extend the number of possible HRCs to beyond a manageable size for this project. Since the error rates and nature (probability distributions) vary significantly across transmission systems, this investigation area may be better deferred to a later effort.

3. Beyond the need for MPEG2 video, we need to have other sets of data (training and testing) for developing objective quality methods in other video quality classes (see 4a and 4b below). The data (ideally) would already be subjectively rated and offered on digital format.

3. Although it may become necessary to divide the tests into two main groups, broadcast (classes 1 - 3) and video conferencing (classes 4 - 6), we should not start with that conclusion. See 2a above.

4. Draft Recommendations on objective assessment of video quality. These Recommendations should eventually cover both in-service and out-of-service objective assessment methods. I feel it is possible to produce Recommendations which initially cover only the areas of most interest to the institutions participating in the process. We can of course work on more than one area at the same time. Please express which areas are of interest to your organization and toward which you are willing to devote resources.

4. Again, we should not make an arbitrary division at this time. We would like to determine one basic technology which provides a reference measurement for all classes if possible. Tektronix is most interested in broadcast (classes 1 - 3) however we would like to participate in the full range of testing to the extent possible based on resources and time.

a. J.OVQ (Covers Classes I, II, and III) (nominally SG9)

i. Class I: Contribution

ii. Class II: Primary Distribution

iii. Class III: Secondary Distribution

4a. Our proposed definition of classes 1 - 3 are based on what will be done with the resulting pictures:

1. Complete post production, many edits and processing layers, intra-plant transmission. Also used for remote site to plant transmission (backhaul).

2. Simple modifications, few edits, character/logo overlays, inter-facility transmission. An example would be network to affiliate transmission.

3. No changes, delivery to the home/consumer.

b. P.OVQ (Covers Classes IV, V, and VI) (nominally SG12)

i. Class IV: All frames encoded. Low Artifacts. Usually 30 fps.

ii. Class V: Frames may be dropped at encoder. Perceptual artifacts possible, but useful for designed tasks.

iii. Class VI: Series of stills. Not intended to provide full motion. (Examples: Surveillance, Graphics).

4b. We agree with the definition for classes 4 - 6.

c. ITU-R WP11E Recommendations

4c. We are willing to work on draft recommendations for ITU-R WP11E