

Submission to G-2.1.6

**Progress Report of Task Force
to define
A Unit of Measure and Means of Calibration
for
Video Quality Analysis**

Author: Leon Stanger
1 Jan 99

Task Force Members:

Alan Godber
Roger Green
Paul Haskell
Paul Jones
John Libert
Bruce Lilly
Al Morton
Mihir Ravel
AnnMarie Rohaly
Leon Stanger
Andrew Watson

Goal:

To define a unit of measure and a calibration method for video quality impairments. The unit of measure must be defined in terms of human perception and calibrated by statistical analysis of viewer test results.

Activity This Period:

Since the last meeting at Kissimmee FL, the task force has focused on a number of discussion points. These are key points of interest that will need to be tied down before subjective testing begins.

There has been a considerable amount of discussion on several points with a great deal of thought and insight. My thanks to all of the committee members for their contributions.

The comments are compiled as follows:

Discussion Points:

DP 1: Two Screens vs One

At the meeting in Kissimmee, we discussed 4 options: 1) Two screens side by side, 2) One screen with sequential video clips, 3) Split screen with side by side images, and 4) One screen with the viewer having a switch to toggle between "A" and "B" at will.

Each method has unique pros and cons. Here are a few of the basic thoughts that were discussed:

1) Pro: Two screens allow a viewer to glance between screens quickly to see subtle differences.

1) Con: It's tough to perfectly match the color, brightness, etc. to prevent the viewer from judging the picture on the wrong things. It is also difficult to overcome left/right preferences.

2) Pro: It's an easy test to administer. If the video segments are kept short and repeated several times, it should be possible for the viewer to retain a mental image long enough to see a difference.

2) Con: Due to the memory factor, one might not be as sensitive to small differences.

3) Pro: A split screen is the best of both worlds of 1) and 2).

3) Con: It is an expensive test. It requires a DVE (digital effects generator) to slide the picture so the viewer can see the same part of "A" and "B" picture. It does not lend itself to man/machine correlation tests. Analysis tools are intended to measure a full frame sequence from a device under test, not a partial image.

4) Pro: It solves the problems in 1) and still maintains the ability to look at small differences.

4) Con: It takes away a degree of lab control.
Leon Stanger

A DVE is not necessary and might not be advisable. A small production switcher with wipe capability is sufficient. To avoid left/right and top/bottom preferences, I propose that a given pair of sequences should be shown several times, with diagonal wipe transitions (both NW/SE and NE/SW). An advantage is that the two images are adjacent, thus avoiding loss of sensitivity to small differences which may result with side-by-side monitors (and also with sequential presentation on a single monitor).

Bruce Lilly

I favor (2).

Option (1) suffers from the difficulty that one cannot look at both screens at once. Highly visible artifacts may go by unnoticed while one is looking at the other screen.

Option (3) suffers from the same defect, and in addition prohibits one from looking at video in its "natural" dimensions.

Option (4) increases variability, and introduces a signal (the transient between reference and test) that is not present in the real-world video

experience.

Andrew B. Watson

I think option (2), successive presentations on a single monitor, will be the most practical method without introducing confounding effects.

John M. Libert

DP 2: Fractional JND

At the Kissimmee meeting, this point was discussed. There was a short discussion between the two alternatives: 1) a different percentage of observers, and 2) a mathematical extrapolation.

Method 1) could arbitrarily define a point such as "1/2 JND is the point where 62.5% of the viewers prefer A to B." Even though it could be statistically measurable, the data becomes less certain as the difference gets very small ie 0.1JND.

Method 2) would only relate to objective measurement algorithms. This method would have no correlation to subjective viewing.

Leon Stanger

DP 3: Multiple Types of Artifacts

At the meeting in Kissimmee, Discussion Point 3 was presented. I believe that we had general agreement that we need to keep the door open for future work on multiple picture defects but in the near term, keep it simple by analyzing one impairment at a time.

Leon Stanger

I am not sure what "one impairment" means, but I think it is best at the outset to deal with real-world artifacts eg MPEG artifacts.

Andrew B. Watson

Beau,

You asked what "one impairment" meant.

The discussion was centered around typical MPEG artifacts. Depending on scene content and other factors, we may see blocking, mosquito noise, or motion artifacts. One impairment meant to develop material which had predominantly one type of artifact rather than a mixture of multiple artifacts.

Leon Stanger

DP 4: Lossless Distribution Format

At the Kissimmee meeting, it seemed that everyone was satisfied that D1 or D5 tape formats were suitable for distribution of test material.

Leon Stanger

Any computer readable format (DAT, DLT, Exabyte, CD, DVD) is ok for us, but

not D1 or D5.

Andrew B. Watson

DP 5: Observers Should Have Corrected Vision

At the Kissimmee meeting DP 5 was discussed. Of course observers should have corrected vision. The issue is what other test parameters should be tied down so that different labs can produce highly correlated results?

It was suggested that we should use Rec 500 as a guideline. Dave Fibush gave me the reference of Alexander Schertz as a contact point.

Leon Stanger

Leon-

I have been working with VQEG to define the vision screening procedures for its upcoming subjective test. The decision was made to follow Rec. 500 guidelines and test for normal visual acuity and normal color vision.

A proposal was also made to conduct a contrast sensitivity screening in addition to the usual Rec. 500 tests. The rationale for including a contrast sensitivity measurement is: (1) human vision-based models of image quality incorporate (and may be calibrated to) contrast sensitivity functions and (2) contrast sensitivity measures are more predictive of performance on "everyday" visual tasks than are visual acuity measures. While a consensus was reached regarding the benefits of such a screening, the contrast sensitivity test was ultimately dropped from the subjective test plan due to the lack of availability of a reliable test chart.

My understanding is that the preferred chart will be back in production soon so it may be possible to include such a screening in our subjective testing.

Ann Marie Rohaly

DP 6: Difference vs Preference

In my Draft 3 requirements document (18 Oct 98) I proposed a preference of sequence A to sequence B. Some of the comments from the group were:

- 1) It's easier to see a difference than determine which is better.
- 2) Preference could lead a person to "like" a picture with more of a particular artifact. Sometimes people like softer pictures or more noise.
- 3) A difference may imply a lower threshold than a preference.
- 4) A difference has the risk of extraneous factors like a subtle color or brightness differences influencing the outcome. Indeed there might be a difference but it might not relate to a blocking artifact that we're trying to gather data on.

A fundamental issue at stake is the notion of a threshold. On one hand, it should be the smallest detectable difference. On the other, it should be a point of low uncertainty that can be easily duplicated in multiple labs. I lean toward the latter. If we cannot get highly correlated results between labs our work will have little credibility.

So which is the better answer? Are there studies that might lead to a preferred method?

Leon Stanger

I believe that difference is the correct parameter which we should be determining.

Preference allows for judgement by the viewer of what they like, and viewers may not agree about that issue, which would lead us to variable results. We also would not know what caused the variable results.

To overcome the problem of unintended differences, which you discuss in 4, we can either inform viewers what differences (type of artifact) they should be looking for and ask them to ignore other differences, or if we wish to gather more information, we could record the thresholds for the desired difference and for the undesired difference, if there is an additional effect(s).

Either way, making sure that the viewer is reporting on the same artifact as other viewers is very necessary to this process which we are proposing. This separation of artifacts is not the way other subjective viewing has been done, in that all artifacts are taken into consideration by the viewer without identifying the individual artifacts when determining a score using Rec. 500 methods.

The method we are proposing, as a result, should have much greater repeatability than current methods.

Highly correlated results between labs and viewers should be a fundamental aim.

Alan Godber

DP 7: Size of Screen

At the Kissimmee meeting, this point was kicked around a bit with viewing distance or direct view vs projection. In the end, I don't think anyone had any objection to a direct view CRT 19 to 21 inches with a 4x3 aspect ratio as the standard for the test.

Leon Stanger

19 - 21 inch pro monitor is fine with me.

John Libert

DP 8: Definition of One JND

Again, this point ties closely to DP 6: Difference vs Preference.

It seems that the simple model is: If "Preference" is used, the 75% point has the lowest uncertainty. If "Difference" is used the 50% point is suitable.

Since the 50% point has 6dB less "noise" than the 75% point, it would suggest some advantage of using difference as the preferred method.

A reference was made to the Sarnoff paper presented last year on a JND. We need to reread that paper.

Leon Stanger

Task Force,

Here is an excerpt from a Sarnoff paper describing their JND model:

"...the JND unit of measure is functionally defined such that 1 JND corresponds to a 75% probability that an observer viewing the two images multiple times would be able to see the difference. JND values above 1 are then calculated incrementally. For example, if Image Y is 1 JND higher in contrast than Image X, and Image Z is 1 JND higher in contrast than Image Y, then Image Z is 2 JNDs higher in contrast than Image X. In probability terms, this 2 JND difference corresponds to 93.75% probability of discrimination ($0.75 + 0.75 \cdot (1 - 0.75)$), and a 3 JND difference corresponds to 98.44% probability. Although probability of discrimination asymptotes quickly as a function of JNDs, the units are useful because they correspond to roughly linear magnitudes of subjective visual difference..."

John Libert

DP 9: 10 JNDs

The issue at stake is how many step and repeat JNDs can subjective testing produce? There wasn't too much discussion on this point since no one seemed to have a good feel for the accuracy beyond 2 or 3 JNDs.

Leon Stanger

DP 10: More Experts

John Libert will attempt to contact the list of references identified by Beau Watson.

Leon Stanger