VIDEO COMPRESSION PERFORMANCE: SUBJECTIVE EVALUATION OF PICTURE QUALITY

Bronwen Lindsay Jones, Independent Contractor
Richard S. Prodan, Ph.D., Cable Television Laboratories, Inc.
David A. Eng, Cable Television Laboratories, Inc.

Abstract

An experimental design incorporating psychophysical test methods for use in the subjective assessment of picture-quality variations due to compression artifacts is described. Standard-definition television pictures are used as test material, including video, film and some extensively utilized MPEG (Moving Picture Experts Group) test sequences.

This study is unique and may be the first of its kind. The effect of compression artifacts on picture quality is very subtle in nature. There are many trade-offs between and among bit rate, compression tools, resolution, picture content and recording format.

A combination of test methods is being utilized, based in part on a new International Telecommunications Union-Radiocommunication Sector (ITU-R) listening test procedure for testing very small differences.

Introduction

The advent of differing compression schemes and bit-rate combinations for video and audio information has caused confusion in the industry (over 100 combinations have been reviewed to date). Optimization or best-fit for viewer preference and expectation of picture quality is expected to vary with program content and demographics. Answers which will help guide industry standards are needed as are recommendations regarding variability with program content and source format.

Because standard-definition, compressed-digital picture artifacts have not been examined in any formal subjective manner, their effect on perceived picture quality to various viewer populations is not known. This proposal describes an initial small pilot study aimed at getting these answers, using industry experts as observers.

Method

The uniqueness of this study is in the new and very subtle nature of the manner in which compression artifacts manifest themselves in pictures. Objects in motion can exhibit blockiness, edge business, and a shimmery, twinkling "mosquito" noise that is visible in high contrast areas around sharp transitions. Test methods must be chosen carefully to fit experimental test conditions in order to avoid overly influencing the outcome. For example, compression artifact differences are small. A subjective scale with too little resolution will show large differences but not small ones. Small differences are perceived, but will not show up in the data.

Due to the expertise of the observers, the small differences in compared picture quality, and the internal nature of the initial study, a special combination of test methods is being employed. Unlabeled graphic scales are presented for recording viewer judgements in a continuous, proportional ratio-scale manner. In addition, viewers are encouraged to record comments in an information gathering technique which has become known in the industry as Expert Observation & Commentary (EO&C).

Observers

Expert observers from within CableLabs, including the authors/experimenters, set up the

study and finalized the choice of appropriate test methods and test materials.

Expert observers from within CableLabs' membership (i.e., industry experts and video engineers) will be used as subjects in these pilot studies; non-expert viewers may be added later for more generalized test results. Experts are defined by the ITU-R (formerly CCIR) as "observers who have had recent extensive experience in observing picture quality or impairments, particularly of the type being studied."

Picture Source Material Selection

New motion-picture test material has recently been assembled by groups such as The Moving Picture Experts Group (MPEG), the Federal Communications Commission's Advisory Committee on Advanced Television Systems Planning Subcommittee Working Party 6 on Subjective Assessments (ACATS PSWP6), and some private parties. It generally consists of 10-to-35-second segments of video material (no audio) originated on film, on HDD 1000, on D1, and on Beta SP. This ensured excellent quality origination. Such quality origination is especially important for source comparisons.

Additional test material originated on film was provided by Lucasfilm Ltd. Test material that was representative of high-motion cable sports programming was provided by ESPN Engineering. Video sequences from the original CableLabs/Viacom compression test source were included.

The choice of which test material selections to use in an experiment is made by expert review and includes seeing all test material under all viewing conditions. Those sequences which are most sensitive to the impairments or artifacts being studied are chosen. The selected sequences are displayed according to system M 525-line component standards, after compression and expansion. As a result, a compilation of 17 selections from MPEG, Lucasfilm Ltd., ESPN, and Viacom are the primary test-material grouping. The material is arranged for presentation in several blocked and balanced pseudorandom orders.

Test Material Production

A studio-quality, Panasonic D-5 digital component VTR input source material to a digital video compression encoder connected to a decoder produced by the same vendor.

The encoder output fixed values of constant bit rate. The output of each decoder was tape recorded with a second studio-quality digital component VTR. A short segment of compressed digital information in MPEG transport stream format was captured and verified by CableLabs. The verification process included bitrate, resolution and profile elements, as well as MPEG-2 bitstream syntax at the system and video layers.

Test Equipment

Hardware included a studio-quality video tape recorder (Panasonic model AJ-D580P), the randomized tapes, and two or more Sony BVM-1911 CRT monitors. If viewer response-gathering becomes automated at a later date, computer keyboards may be included.

The study was conducted in a viewing environment closely matched to ITU-R viewing-room specification. One to three viewers participated at a time.

Experimental Test Design and Procedure

This study was segmented into two or three primary blocks:

1. Source and four compression techniques including full CCIR-601 resolution MPEG Main Profile (MP) using I, P, B frames; Simple Profile (SP) using I, P frames with MPEG-2 dual prime smart prediction; intra-coded and predicted-frame coded (IP)

MPEG-2 profile using I, P frames without smart prediction; and a proprietary profile using I, P frames with DigiCipher™ extensions. Processing was accomplished by a single manufacturer's encoder/decoder at bit rates of 8, 6, 4.5 and 3 Mb/s;

2. Source and full resolution MPEG MP, SP, IP, and proprietary processing as generated for comparison by two primary encoder/decoders at bit rates of 8, 6, 4.5 and 3 Mb/s;

3. Source, full, and 3/4 resolution MPEG MP, SP, IP and proprietary processing as generated by all encoder/decoders at bit rates of 8, 6, 4.5 and 3 Mb/s.

Block One is an evaluation of a single coding system operating at full resolution, but at different bit rates and with different profiles to determine just exactly how similar they are. There are as many as 32 possible operating conditions tested. Block two and three evaluate between and across manufacturer's systems.

Side-by-side pair comparisons, in parallel rather than in sequence, will be the primary response-gathering procedure. Unlabeled graphic scales will be used in a continuous, proportional, rather than discrete, manner. (This approach asks not only if something is better, but how much better; rating scales ask merely for a position on a continuum). Both members of each pair are judged and scored, in accordance with ITU-R Recommendation BS 1116. Expert Observation & Commentary (EO&C), as was used in the listening tests for the FCC ACATS in the US audio standards effort (recently concluded end of 1995, HDTV sound) will accompany the scaling procedure.

Test material will be made up of different groupings of the 17 test material selections previously described. The viewing distance from the screen (three or four times the picture height) was chosen by the expert observers and noted and reported.

Block one may be conducted repeatedly with any or all systems if deemed in need of a similar thorough evaluation. Blocks two and three will be used in the same manner for comparison across systems: ratio scales in side-by-side pair comparisons using EO&C.

Depending on the visibility of artifacts, it may be necessary to impose top and/or bottom anchors by including uncompressed source and/or very low bit rate compressed test material.

Presentation of Results

The presentation of results includes graphs and tables of means and standard deviations. Ratio scaling of data usually makes use of geometric, rather than authentic, means and standard deviations in order to lessen the influence of the spread of the members (proportions are of interest, not numerical values).

The size of this study is fairly small and manageable, therefore, viewer responses can be gathered with paper and pencil and the analysis of results can be done manually. If demonstrations become important or the number of observers grows a great deal, it may become desirable to automate viewer-response gathering. When this is done, the statistical processes are incorporated into the software and data can be plotted immediately upon completion of the test. This kind of automation also considerably speeds up data analysis and report writing.

References

1. Recommendation ITU-R BS.116, Methods for the Subjective Assessment of Small Impairments in Audio Systems Including Multichannel Sound Systems.

2. Recommendation ITU-R BT.500-6, Methodology for the Subjective Assessment of the Quality of Television Pictures.