CCITT SGXV
Woking Party XV/1
Experts Group for ATM Video Coding

AVC-176

INTERNATIONAL ORGANIZATION FOR STANDARDIZATION
ORGANISATION INTERNATIONALE DE NORMALISATION
ISO/IEC JTC1/SC2/WG11
CODING OF MOVING PICTURES AND ASSOCIATED AUDIO

ISO/IEC JTC1/SC2/WG11
MPEG91/336
NOVEMBER 1991

Source: Bellcore/NTT
Title: Picture Performance of 1.3, 4 and 9 mb/s Bellcore Algorithms and Other Test Processes
Purpose: Information Contribution
Subgroup: Video

1. INTRODUCTION

A subjective study has recently been completed at Bellcore to evaluate the picture quality associated with compressed bit rate algorithms operating at 1.3, 4 and 9 mb/s, and the picture quality of other benchmark test processes. The purpose of this contribution is to present and discuss the subjective test plan and the results of the tests.

The test processes included three compressed bit rate encoding algorithms, one operating at 1.3 mb/s, and the other two operating at both 4 and 9 mb/s and three other processes including the unprocessed source pictures of CCIR 601 quality, an NTSC process and a low-quality "dummy" process. The source pictures were the eight picture sequences planned for use in the MPEG 2 assessment[1] [2]. The NTSC process was obtained by recording the source pictures on a 1 inch video tape from a C-type video recorder. The dummy process was obtained by making tandem recordings of the source pictures on a commercial VHS VCR.

The 1.3 mb/s algorithm is Bellcore's latest version of the MPEG 1 algorithm and is referred to in this contribution as Bellcore'91.

One of the algorithms operating at 4 and 9 mb/s is MPEG 1 compatible and the other is not. The subband hierarchical algorithm, referred to as SB4 or SB9, operates by first decomposing the input picture into 4 subbands. The lowest frequency picture is encoded using a pyramidal coder that generates an MPEG 1 compatible bit stream as well as additional information that improves the quality of the low-band image. The remaining bands are also encoded using MPEG 1 compatible coders. The resulting coder is forward and backward compatible with MPEG 1.

The extended MPEG algorithm, referred to as MPEG+4 or MPEG+9, is similar to the MPEG 1 algorithm running at a higher rate. However, in order to improve picture quality a number of changes were made in the MPEG 1 algorithm. A new motion compensation scheme is used to accommodate interlaced inputs (MPEG 1 takes only progressive inputs); a better buffer control strategy is used that takes the picture content into account when allocating bits; and the variable length codes are dynamically adapted to keep code efficiency high.

The next section describes the test plan, including test conditions, viewing conditions, subjects, and test procedures. Section 3 discusses results and Section 4 states conclusions.

2. TEST PLAN

In general, the test procedures and guidelines given in CCIR Recommendation 500-3[3] were followed. The basic approach used in the subjective test plan was to have nonexpert observers rate the quality of pairs of 5-second video sequences representing various test processes. They used a continuous quality scale that ranged from EXCELLENT to UNSATISFACTORY. Details are discussed in subsequent subsections.

2.1 Test Pictures

The eight, 5-second picture sequences that were selected for the MPEG 2 assessment[1][2] were used in the tests. The pictures are entitled "Flower Garden", "Mobile & Calendar", "Suzie", "Tempete", "Football", "Popple", "Table Tennis", and "Edited". All of these were originally obtained on 8 mm exabyte tape from the MPEG test sequence distributor and then transferred to Dl tape for test material preparation.

2.2 Test Processes

The following eight processes were tested: (1) Source, (2) NTSC, (3) Dummy, (4) Bellcore'91, (5) MPEG+4, (6) MPEG+9, (7) SB4, and (8) SB9.

The source process was represented by the CCIR 601 quality, MPEG 2 picture sequences.

The NTSC process was the result of converting the RGB signal output from the source sequences on the Dl tape to NTSC and then recording it on a 1 inch video tape. This process resulted in a slight color shift in the NTSC pictures that was not noticed until after the subjective tests had begun. Although it is not certain how much effect, if any, this color shift had on the NTSC results, it is believed that the effect was minimal.

The low-quality dummy process was included to ensure full coverage of the rating scale. It was produced by converting the RGB signal output from the source sequences on the Dl tape to NTSC and then making tandem recordings on a commercial VHS VCR.

The remaining 5 processes were the compressed bit rate algorithms as described in the Introduction-Bellcore'91 at 1.3 mb/s, MPEG+4 and SB4 at 4 mb/s, and MPEG+9 and SB9 at 9mb/s. Bellcore'91 is Bellcore's latest version of the MPEG 1 algorithm. The SB4 and SB9 algorithms are forward and backward compatible with MPEG 1 but the MPEG+4 and MPEG+9 algorithms are not.

The 8 test pictures in combination with the 8 test processes resulted in 64 test sequences which were stored on a Dl, digital source tape. Test tapes were then produced by recording the appropriate test sequences from the Dl source tape onto Dl test tapes.

2.3 Test Method

A double stimulus test method using a pair of continuous quality rating scales was used in the tests. In this method each test condition consists of a pair of test pictures sequences, one of which is always the unimpaired reference picture. The reference was the unprocessed, CCIR 601 video which was also used as a test process and, therefore, paired with itself in some of the test conditions. The position of the reference picture was changed in a random fashion and the observers were not told that one of the test picture sequences was a reference. The effect of order was tested by repeating each of the test conditions and reversing the order of the reference on the second presentation. The observers rated each of the picture sequences individually by marking a continuous quality rating scale.

Figure 1 shows an example of the pair of rating scales used for each test condition. Each scale was divided into 5 equal areas with corresponding quality terms of EXCELLENT, GOOD, FAIR, POOR and UNSATISFACTORY. The subjects could place a mark anywhere on the scale that they judged corresponded to the picture quality.

A total of 29 nonexpert subjects, selected from an out-of-house subject pool, participated in the tests. Three were male and 26 were female. Their ages ranged from 28 to 66 with an average age of 51. Each was required to have normal vision acuity (equal to or better than 20/25) and normal color vision.

2.3.1 Viewing Conditions A 19-inch Conrac, 6545 monitor was used in the tests. The monitor controls were adjusted to provide a peak luminance of approximately 23 footlamberts and a minimum luminance of approximately 0.15 footlamberts for a contrast ratio of about 150. The screen color temperature was set at D6500.

The test room was approximately 12 x 20 feet with light grey, sound absorbing walls and controlled, incandescent ceiling lighting. The ambient lighting in the test rooms was dimmed to approximately 0.6 footcandles (fc) at the subject seating positions by adjusting the dimmer controls for the ceiling lamps. The resulting ambient lighting at the monitor was approximately 0.1 fc.

Three observers were tested at the same time in each of nine tests and two observers were tested in a tenth test. (One failed to appear as scheduled.) The observers were seated in front of the monitor at a viewing distance of 5 times picture height (5H). The 5H distance is in the middle of the range recommended in CCIR Rec. 500-3.

2.4 Test Procedures

Eight test processes and eight picture sequences resulted in 64 basic test conditions. Each condition took 24 seconds and consisted of two 5-second picture sequences, two 5-second rating periods and two 2-second message screens identifying each particular sequence. Each condition was repeated twice with the position of the reference reversed for a total of 128 test conditions.

The subjects were seated, three at a time, in front of the monitor at a viewing distance of 5H. They filled out their name and indicated their seating position on the information page of a notebook which contained all of the rating scales for the practice and test conditions. The test administrator then read the test instructions. After answering any questions, a practice session consisting of 8 conditions was conducted. Any problems or questions resulting from the practice session were then resolved before starting the actual test.

The test was administered in two parts with a half-hour break between the parts. The first part of the test consisted of all 64 test conditions presented in a random order. The second part consisted of a repeat of the 64 test conditions but with a different random order and with the position of the reference reversed from where it was in the first part. The random orders were changed for each group of subjects. (The random orders were not truly random as the same picture sequence was not shown twice in a row as recommended in CCIR Rec. 500-3.)

3. RESULTS

3.1 Data Analysis

The first step in data analysis was to convert the marks placed by the observers on the rating scales to numerical scores. This was done by using a transparent overlay that divided the vertical scale into 100 units with 0 corresponding to the poorest quality and 100 to the highest quality. The difference between the numerical scores for the reference and test sequences was calculated for each test condition. Average numerical scores and average differences were then calculated across various cross-sections of the data. An Analysis of Variance (ANOVA) using a mixed factors model[4] [5] was also performed on the difference data.

3.2 ANOVA Results

The major factors in the ANOVA model were the 8 test processes, the 8 picture sequences and presentation order (reference/test, test/reference). The effects associated with processes and picture sequences were highly significant (probability of the effects occurring by chance less than .0001). However, the presentation order did not produce any significant changes in ratings.

3.3 Average Differences

Figure 2 shows the differences in ratings between the reference (unprocessed, CCIR 601 source sequences) and test processes, averaged across picture sequences, observers and presentation order. The average for the NTSC test process includes data from only 26 subjects. In producing one of the test tapes, a band of random noise was inadvertently added to the NTSC picture sequences-the data from these conditions was, therefore, discarded. A difference of 20 corresponds to one comment area of the five on the continuous rating scale (the higher the difference, the poorer the quality of the test process relative to the reference). For clarity, the results for the dummy test process are not shown. The average difference between it and the reference was 67.7.

The average difference for the conditions pairing the source (reference) against itself was -0.14, very close to the expected difference of 0. The average difference for the NTSC test process was 10.8. As stated earlier, a slight shift in color occurred in the production of the NTSC test sequences. Although not known with certainty, it is believed that this color shift had a minimal effect on results.

The average differences for the MPEG+4 and SB4 compressed bit rate algorithms were 5.8 and 10.8, respectively. The average differences for the MPEG+9 and SB9 algorithms were 2.8 and 5.6, respectively. A t-test for related measures was performed to determine the statistical significance of the difference in the means between the MPEG+ and SB algorithms at both 4 and 9 mb/s. In both cases the differences were highly significant (probability that differences occurred by chance variation much less than .0001). These results indicate that the MPEG+ compression algorithm produces higher picture quality than the SB compression algorithm. The SB algorithm at 4 mb/s appears to be roughly equivalent in picture quality to the NTSC process. The picture quality of the MPEG+ algorithm at 4 mb/s and the SB algorithm at 9 mb/s appears to be about half-way between that of NTSC and CCIR 601. The picture quality of the MPEG+ algorithm at 9 mb/s appears to approach that associated with CCIR 601.

The average difference for the 1.3 mb/s algorithm, Bellcore'91, was 26.5. This difference is substantially higher than the differences for the 4 and 9 mb/s algorithms and for the NTSC process and, thus, indicates a lower picture quality.

The ANOVA indicated that effects associated with picture sequence were also significant. Figure 3 shows the average difference ratings for each of the picture sequences for each of the test processes (for clarity, results for dummy process are not shown). Clearly, subjective reaction depended on the particular picture sequence being tested. The most sensitive sequences to the NTSC process were "Mobiles & Calendar" and "Edited". All of the sequences were sensitive to the 1.3 mb/s, Bellcore'91 algorithm, but the most sensitive were "Mobiles & Calendar" and "Football". The sequences "Flower Garden" and "Popple" were most sensitive to the MPEG+ algorithms while "Mobiles & Calendar" and "Football" were most sensitive to the SB algorithms. In general, the sequences "Susie" and "Tempete" were least sensitive to the various test processes.

4. CONCLUSIONS

Results of this study indicate that Bellcore's extended MPEG (MPEG+) compressed bit rate algorithm produces higher picture quality at both 4 and 9 mb/s than does Bellcore's subband hierarchical (SB) compressed bit rate algorithm. Quality differences between MPEG+ and SB were highly significant. The SB algorithm at 4 mb/s is roughly equivalent to NTSC picture quality. The picture quality of the MPEG+ algorithm at 4 mb/s and the SB algorithm at 9 mb/s is about half-way between that of NTSC and CCIR 601. The picture quality of the MPEG+ algorithm at 9 mb/s approaches CCIR 601 picture quality. Picture quality of Bellcore's 1.3 mb/s, MPEG 1 compatible algorithm is lower than that of the MPEG+ and SB algorithms and of the NTSC test process.

REFERENCES

1. ISO/IEC JTC1/SC2/WG11, MPEG 91/025, "D-1 Tape Editing for MPEG-2 Assessment," Tsuneyoshi Hidaka (JVC), May, 1991.
2. ISO/IEC JTC1/SC2/WG11, MPEG 91/019, "Test Sequences for MPEG 2," Dietmar Hepper (TCE/DTB), May, 1991.
3. CCIR Recommendation 500-3, "Method for the Subjective Assessment of the Quality of Television Pictures," 1986.
4. Roger E. Kirk, "*Experimental Design Procedures for the Behavioral Sciences,*" 1968, Wadsworth Publishing Co., Inc..
5. Leonard A. Marascuilo and Ronald C. Serlin, "*Statistical Methods for the Social and Behavioral Sciences,*" 1988, W. H. Freeman and Company.

FIGURE 1 RATING SCALES USED IN TESTS

TEST CONDITION________

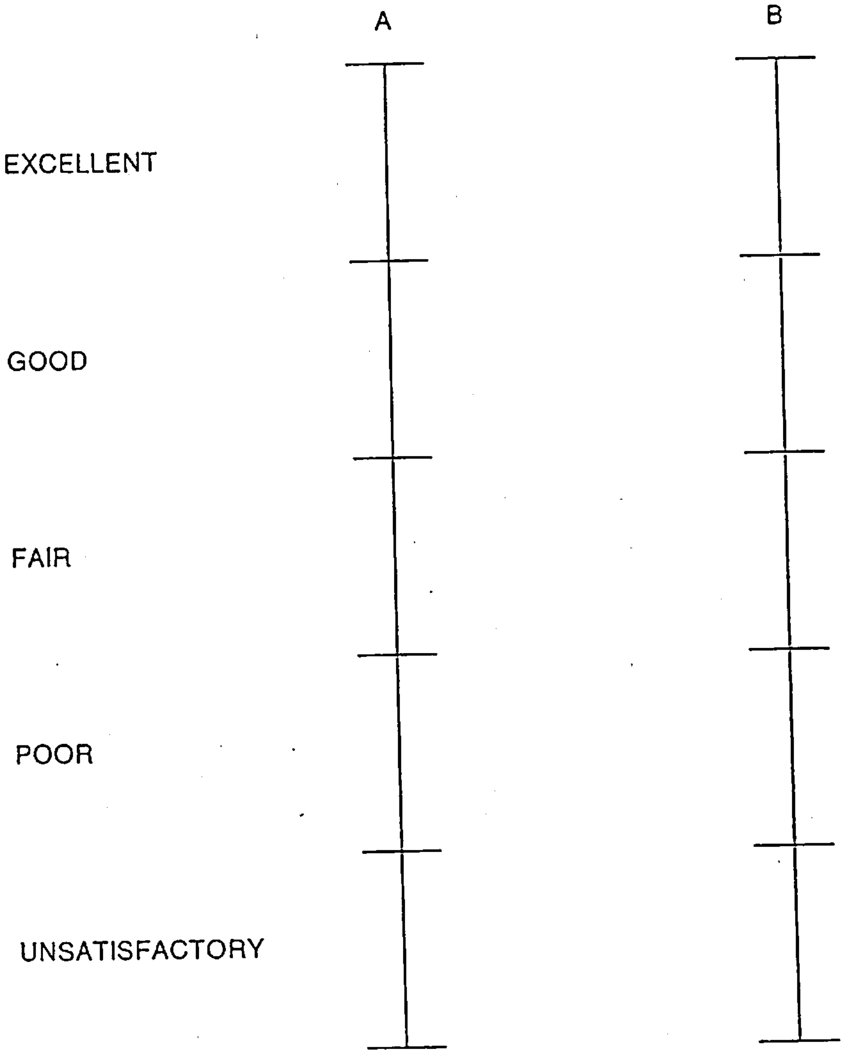


FIGURE 2 AVERAGE RATING DIFFERENCES FOR TEST PROCESSES

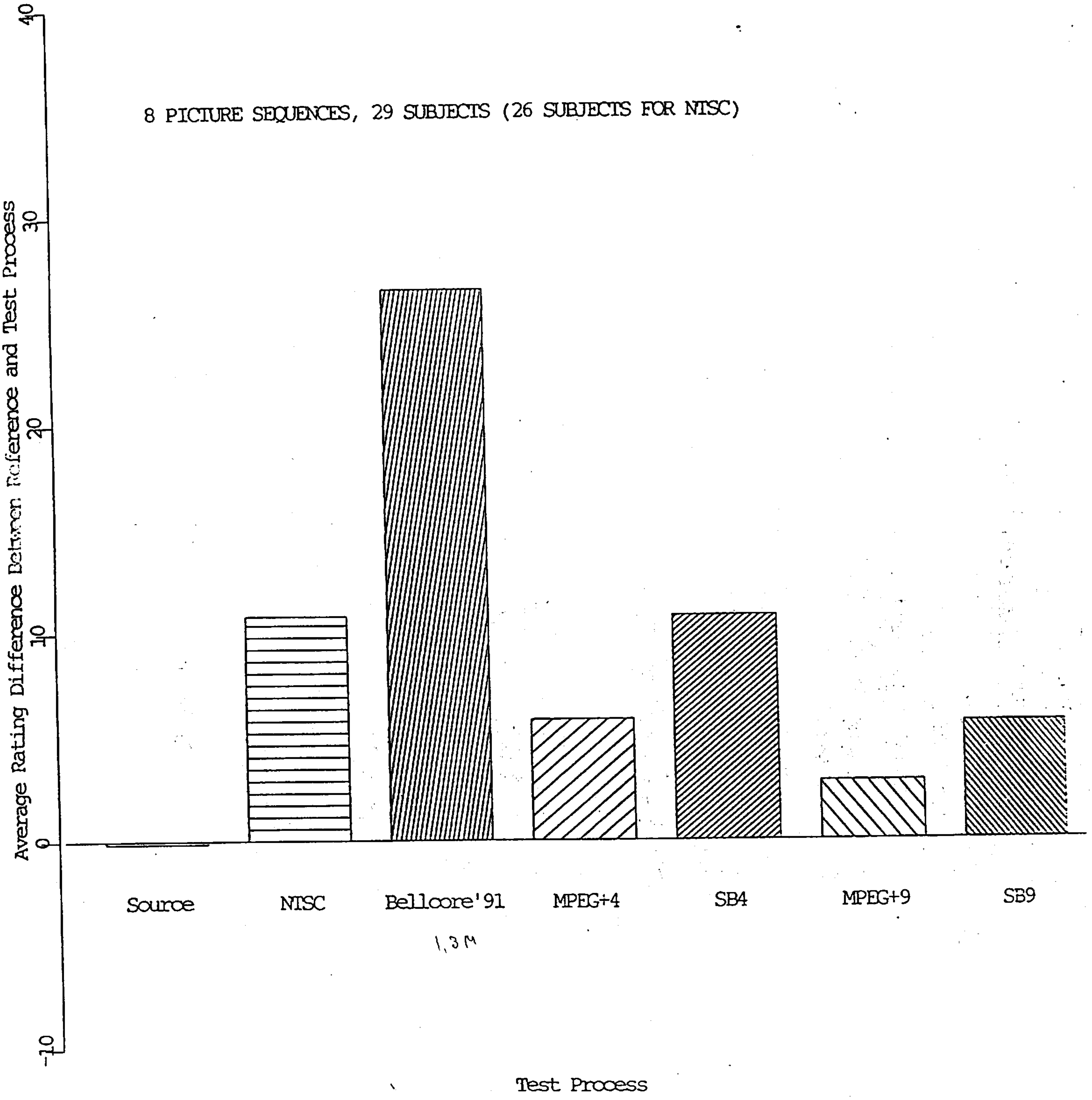
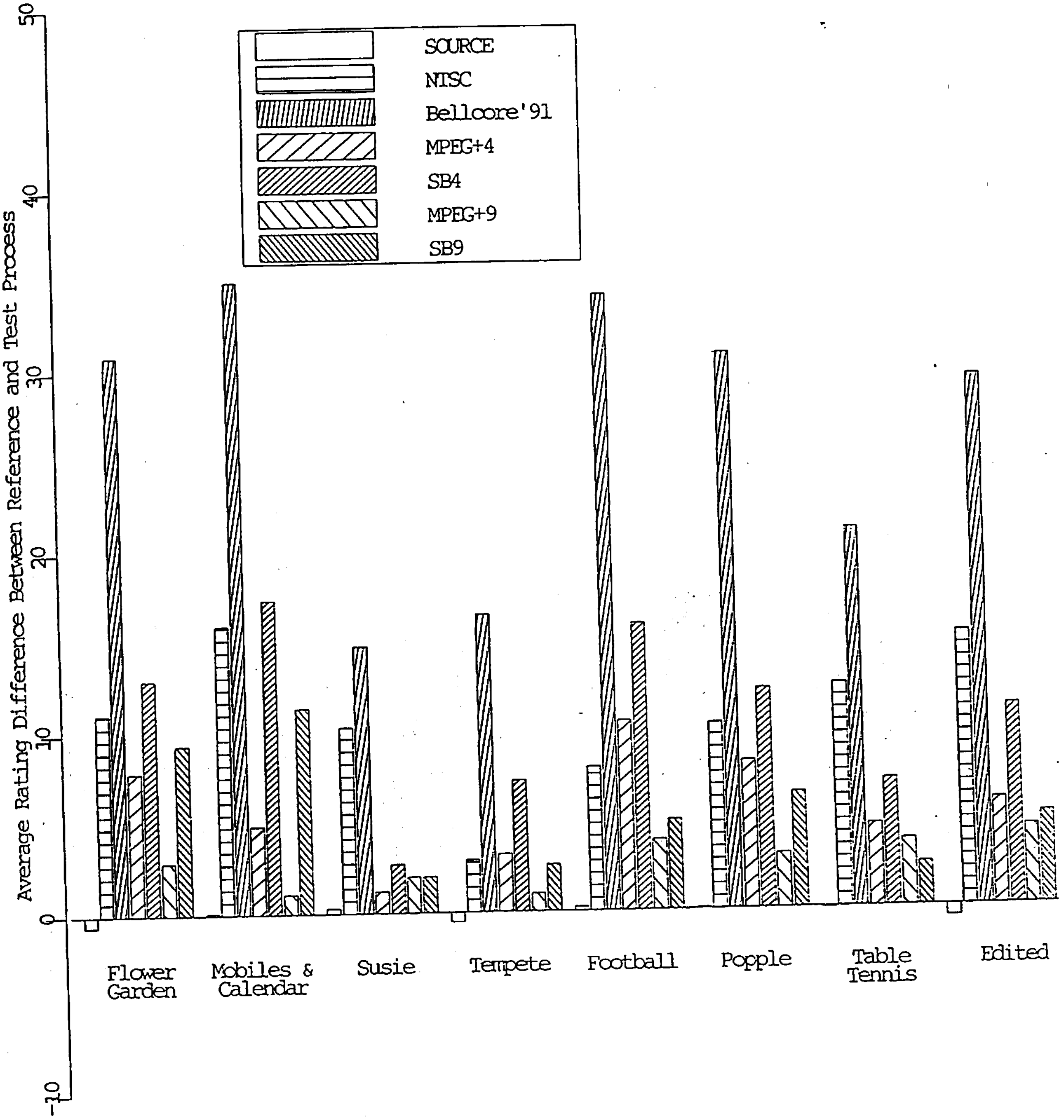


FIGURE 3 AVERAGE RATING DIFFERENCES FOR TEST PROCESSES/PICTURE SEQUENCES


SOURCE
NTSC
Bellcore'91
MPEG+4
SB4
MPEG+9
SB9
Average Rating Difference Between Reference and Test Process
Flower Garden
Mobiles & Calendar
Susie
Tempete
Football
Popple
Table Tennis
Edited
Picture Sequence