NTIA Technical Memorandum 18-532

ITS4S: A Video Quality Dataset with Four-Second Unrepeated Scenes

Margaret H. Pinson

technical memorandum

U.S. DEPARTMENT OF COMMERCE • National Telecommunications and Information Administration

NTIA Technical Memorandum 18-532

ITS4S: A Video Quality Dataset with Four-Second Unrepeated Scenes

Margaret H. Pinson

U.S. DEPARTMENT OF COMMERCE

February 2018

DISCLAIMER

Certain products, technologies, and corporations are mentioned in this report to describe aspects of the ways that images, videos and cameras are used at present or may be used in the future. The mention of such entities should not be construed as any endorsement, approval, recommendation, prediction of success, or that they are in any way superior to or more noteworthy than similar entities that were not mentioned.

CONTENTS

FIGURES

TABLES

ABBREVIATIONS/ACRONYMS

ACR	absolute category rating
FR	full reference
HDTV	high definition television
HRC	hypothetical reference circuit
INLSA	iterated nested least squares algorithm
ITS	Institute for Telecommunication Sciences
MOS	mean opinion score
NR	no reference
PVS	processed video sequence
RR	reduced reference
RSRC	related source sequence
SRC	source video sequence
VQEG	Video Quality Experts Group

ITS4S: A VIDEO QUALITY DATASET WITH FOUR-SECOND UNREPEATED SCENES

Margaret H Pinson[1]

This report describes the video quality subjective test **its4s**, including the experiment design and footage attribution. Subjective experiment **its4s** includes 813 unique video sequences, each four seconds in duration. No video sequences are repeated. The goals are (1) to provide insights into the optimal experiment designs for training no-reference (NR) metrics, and (2) to understand the impact of original video quality on mean opinion scores (MOS). Together these goals support the larger goal of progressing research on effective NR metrics. The dataset is freely available for research and development purposes.

Keywords: image quality, subjective testing, video quality

1. INTRODUCTION

Subjective video quality dataset **its4s** was designed to provide insights into improved experiment designs for training no-reference (NR) video quality metrics. Dataset **its4s** emphasizes videos produced by professional videographers. The dataset emphasizes original footage (i.e., as provided by the videographer) to encourage NR metrics that accurately track the quality of such original videos. This report is intended to provide a full and complete description of the subjective test design and implementation. The **its4s** dataset (videos and ratings) has been made available on the Consumer Digital Video Library (CDVL, www.cdvl.org).

Industry has expressed an urgent need for NR video quality metrics. These metrics predict the quality of a video sequence based only on the sequence itself, with no side information. This has proven to be very challenging. Video quality experts have concluded that the existing models are highly inaccurate. In private communications, several industry associates expressed extreme disappointment after running the available NR metrics on their own videos. The validity of those claims is not verified in this report. Our purpose is not to evaluate existing NR metrics, but rather to encourage innovation.

The **its4s** dataset focuses on two factors. First, the metric performance must degrade gracefully in response to new content (i.e., subject matter, camera, editing). Second, the metric must accurately predict the quality of original videos (e.g., broadcast quality, contribution quality, professional cameras, prosumer cameras). Basically, the NR metric must accurately predict the quality of video sequences that do not contain coding artifacts. To address these needs, **its4s** contains 813 unique video sequences, 35% of which contain no compression artifacts. The

[1] The author is with the Institute for Telecommunication Sciences, National Telecommunications and Information Administration, U.S. Department of Commerce, Boulder, CO 80305.

remaining 65% contain simple impairments, to minimize the confounding factor of coding impairments on the original video's quality as the coding bitrate falls.

2. THEORETICAL EXAMINATION OF NR METRICS

The Institute for Telecommunication Sciences (ITS) performed internal investigations into NR metrics with a goal of understanding why this type of metric is so difficult to develop. This section presents our theory, along with proposed steps to address the problems identified. This section provides a high-level overview and makes broad generalizations. Experts may identify special cases that contradict the unproven theory presented herein.

Let us begin by examining the conventional experiment design. Subjective tests are typically designed to include a full-factorial matrix of source video sequences (SRCs) and test conditions, which we will refer to as hypothetical reference circuits (HRC). Fundamentally, the experiment measures whether or not subjects can perceive a difference between two versions of the same stimulus. This design is useful for comparing codecs and bitrates. Most video quality datasets adhere to the conventional *SRC* × *HRC* experiment design and contain video sequences that are around 8 to 15 seconds in duration. The experiment will typically contain many more impairments than scenes. For example, the Video Quality Experts Group (VQEG) HDTV tests are designed around a full matrix of 9 *SRC* × 16 *HRC* [1].

Video quality metric research relies upon second-hand datasets, by which we mean subjective datasets that were designed for other purposes. These datasets are typically designed to compare codecs, bitrates, and network errors. Such datasets do a good job at characterizing relationships among these impairments—but a bad job of characterizing original video sequences and variables that impact their quality. Videos with production quality problems are intentionally omitted. For example, VQEG validation tests specify that all source videos must have good quality or better as evaluated by an expert. Double stimulus rating scales intentionally remove the impact of the original production quality (see [2] or ITU-T Rec. P.913). Single stimulus rating scales like Absolute Category Rating (ACR) evaluate the original production quality, however each experiment typically includes a very small number of SRC, sometimes as few as one or two SRC. Also, we are concerned that scene reuse may impact how people evaluate and rate videos (see [3]).

When multiple subjective datasets are available, each adhering to the conventional *SRC* × *HRC* experiment design, then it is straight forward to merge those datasets into a single dataset. The technique preferred by ITS is the Iterated Nested Least-Squares Algorithm (INLSA) [4]. Basically, INLSA uses full reference (FR) video quality metrics to put subjective video quality data on a single scale. Datasets can also be merged using techniques based solely on overlapping subjective data, as explained in [5], but such subjective data is seldom available. Both methods create merged super-datasets that provide us with a bigger picture and accurately characterize many different impairments.

These second-hand datasets and the super-dataset merging algorithms are well suited for FR metric research and validation. FR video quality metrics (e.g., [6]) compare the reference video to an impaired version of the video. Because a reference video is available for comparison purposes, the metric focuses on characteristics of coding and transmission impairments (e.g., noise, blocking, blurring, jerky motion). The quality of the reference video is taken as a given, and the purpose of the model is to predict deviations from that quality level that result from coding and transmission.

Unexpected problems arise when we move from FR metrics to NR metrics. By definition, NR metrics cannot refer to a reference video to learn how the video was supposed to look. Instead, the NR metric must understand the subject matter of the scene. This results in a shift of focus from characteristics of the coding impairments to the characteristics of the scenes (e.g., coding complexity, coding bitrate, resolution, aesthetics).

Of these, aesthetics is perhaps the most unconventional. We can envision an NR metric that ignores aesthetics and only predicts the quality impact of imperfect camera electronics. The problem is that the end user is unlikely to similarly discount aesthetics in their independent evaluation of the metric. The Video Quality Experts Group (VQEG) has always validated the performance of NR metrics against Mean Opinion Scores (MOS) calculated from Absolute Category Rating (ACR) subjective tests. These subjects are instructed to ignore aesthetics, but they don't do it. Original video sequences depicting appealing content, like pretty women, receive higher MOSs than videos depicting less appealing content. Basically, an NR metric must accurately predict the quality of original video sequences, including videos where impairments come from an imperfect camera or poor videography.

Video content is heterogeneous. There is a huge variety among scenes, and a wealth of possible subject matters and camera responses to those subjects. Before we can hope to develop an NR metric that tracks aesthetics, camera response, and videography skill, we must have datasets that characterize the large problem space described by the enormous variety of video content. Second-hand datasets are unlikely to address this need.

By contrast, the visual impairments from all modern codecs are homogeneous. Modern codecs are closely related as are the underlying mathematics. The bitrate-to-quality response curves of MPEG-2, H.264 (AVC) and H.265 (HEVC) differ, but all produce visually similar coding impairments. Each divides the video into groups of pictures (GOP) that are fairly short (e.g., ¼ s, ½ s, 3 s), and scene cuts typically occur between GOPs. Each codec divides the GOP into blocks with uniform motion and texture. All use the discrete cosine transform (DCT) or a similar transformation. Each codec contains a complicated nest of algorithms, but in the end each block of video is encoded by a subset of the available algorithmic subroutines.

When the MPEG committee evaluates algorithms, they seek video content that has no scene cuts and relatively uniform content. That is, each frame has consistent amounts of spatial detail, edge strength, motion, etc. The goal is that the entire video will trigger similar algorithm responses. The conventional experiment design is well suited to evaluating these regular differences among coding impairments. Like a codec engineer, the MPEG committee wants metrics that will help them optimize the codec behavior on a very small scale, like 64 pixels by 64 pixels by two frames.[2]

More to the point, every method we have discussed so far (e.g., second-hand datasets, the conventional experiment design, technique for merging multiple datasets, INLSA, and FR metrics) focuses on accurately characterizing impairments. They all do a worse job of accurately characterizing camera response and more elusive traits like artistry, aesthetics, and videography

[2] Information about MPEG committee motivation obtained by private communication with Vittorio Baroncini, GBTech, a member of the MPEG committee and a member of VQEG.

skill. That is a problem for NR metric development. The small number of SRCs provides insufficient information about original scene quality to combine these datasets, when the goal of the super-dataset is to accurately characterize aesthetics, camera response, and videography skill. We cannot build the "big picture" from the few SRCs in second-hand datasets.

Recent advances in video transmission also impact NR model development. Packet loss had been a major topic of interest among subjective video quality researchers as a way to understand the impact of network problems on video delivery. Thus, second-hand datasets tend to investigate a variety of bitrates and packet loss rates.

However, video delivery has changed dramatically as adaptive streaming has become popular. Commercial adaptive streaming services contain proprietary heuristics that choose an appropriate resolution for each encoding bitrate. These systems adjust the streaming bitrate as the network becomes congested, so users do not see impairments from packet loss. Basically, the video is divided into short segments (e.g., 10 s duration), and each segment is encoded at multiple bitrates. In addition to the immediate quality response of the current GOP, the changes in network conditions (and thus bitrate and coding parameters) impose a longer term quality response (e.g., that tracks a user's experience over 5 min). With the arrival of adaptive streaming, packet loss is no longer of interest, but the relationship between video resolution and coding bitrate is much more interesting. Older datasets do not reflect this change in priorities.

Second-hand datasets also make assumptions about the duration of a video sequence that is of interest. Temporal integration has long been a point of contention among video quality experts. There are valid reasons why some experts believe that video sequences with scene cuts should be included in subjective tests, and equally valid reasons why some experts believe that video sequences with scene cuts must never be included in subjective tests. Regardless, even 8 to 10 s video sequences without scene cuts are likely to exhibit dynamic quality changes over time and space, unless extreme care is taken when choosing SRCs. The variety of temporal integration functions within published FR metrics indicates that the true underlying temporal integration is more complex than a simple average over time, even when the question is limited to relatively short durations like 8 to 10 s. For example, the NTIA General Model (commonly referred to as "VQM" [7]) uses three temporal integration functions: average, tenth percentile point, and standard deviation.

Older datasets tend to use 8 to 15 s duration videos, which are well-suited to evaluate the impact of either coding quality or packet loss. However, very long video sequences are needed to understand the impact of network problems on adaptive streaming. Arguably, these second-hand datasets are equally ill-suited for NR metric development. They focus on characterizing temporal integration functions that characterize how video quality changes over time.

Theoretically, temporal integration can be treated as a separate problem from the immediate quality impression of a very short video segment. If we had an NR metric that accurately predicted the immediate quality response, that NR metric could feed into a different algorithm that predicts the longer term quality response as the quality changes over time, in response to different characteristics of the scene or changing network conditions.

This leaves us with a dichotomy. Most subjective datasets were designed before adaptive streaming and contain 8 to 15 second video sequences. The set of source videos is small (e.g., 2 to 10 scenes) and likely contains scene cuts and changes in spatial and temporal information levels (e.g., fairly still then fast movement). The datasets and tests were designed to accurately evaluate bitrates, codecs, and packet loss, using a minimal set of video content.

But the goal of an NR metric is to characterize modern video coding systems and source videos (i.e., straight from the camera). Video delivery systems have changed, so the older subjective datasets are less valuable. To understand video codecs in this new environment, we need to focus on the needs of the codec engineer. That means very short videos (e.g., ¼ s to 3 s duration) with no scene cuts and relatively uniform content (i.e., as per the MPEG committee's scene selection criteria). We also need to fairly characterize the heterogeneous wealth of source videos. It is less important to characterize the full range of coding impairments, because the visual responses to these impairments are likely to be similar.

Our theory is that NR metric development is hindered by the use of datasets produced by the conventional experiment design. We believe NR metric development will be aided by subjective video experiments designed to:

- Contain a huge variety of original videos
- Include low quality original videos (e.g., poor aesthetics, camera problems, amateur videography)
- Use the minimum possible sequence duration
- Describe a state-of-the-art video application
- Exclude temporal integration

Dataset **its4s** was designed to test this theory. The **its4s** dataset investigates adaptive streaming of high definition video to a mobile device, because this is an increasingly popular application. Another goal was to determine whether these and other non-conventional design elements would cause problems. Principally, a few subjects might have difficulty rating all sequences, or many subjects might have difficulty rating a particular sequence. Forced choice rating scales hide these two problems (i.e., the subject is not allowed to say "I cannot decide").

3. EXPERIMENT DESIGN

The **its4s** dataset characterizes a generic adaptive streaming system for high definition mobile devices. The experiment design contains some unusual choices, because one goal is to evaluate a new type of experiment design. The core idea is a video quality subjective experiment design that does not reuse scenes. Although we referred to this type of experiment design as "novel" in [3], this paper uses the term "unrepeated scenes."

The video content was drawn from a pool of high definition television (HDTV) and 4K video recorded in a variety of resolutions and frame rates. The "original videos" (as presented to subjects) have been format converted to 720p 24fps (1280 pixels × 720 line). The 24fps frame rate was chosen as it represents a popular frame rate for adaptive streaming services today. The 720p resolution was chosen as it represents a reasonable compromise resolution that is used among a variety of different monitors (e.g., phone, tablet, laptop, television). Our prior experience developing reduced reference (RR) video quality metrics indicates that any robust metric developed on 720p 24fps content can be easily adapted to other resolutions and frame rates. Additionally, first responders identified 720p as the most desirable compromise between resolution and storage, during the interviews summarized in [8].

Thus, the definition of the original (aka source) videos in this test differs somewhat from the usual definition. The original videos in **its4s** contain a variety of format conversion impairments, depending upon the original video format. Some of the original videos were intentionally chosen because the quality (before format conversion) was fair or worse. These videos constitute a minority of the overall experiment (24 of 813 sequences), because their inclusion directly contradicts all prior advice from experts in the field. Still, these low quality original videos are critical to fully characterize the heterogeneous space of all original videos.

The **its4s** dataset uses the related source sequence (RSRC) experiment design described in [3]. Basically, each source video is expanded into a set of related content with similar characteristics.

The source content was drawn from a large variety of video material. All content was edited into 4 s sequences. Some sequences are a few frames shorter than 4 s. Some sequences contain minor editing errors consisting of two or three frames from a prior content at the beginning of the sequence. These editing errors are noted in Table B-11, in Appendix B. There are no scene cuts in any of the sequences.

The scene duration of 4 s was chosen because this was the shortest sequence duration that could be comfortably viewed and rated by our prototype subjects. Private communications with Philip Corriveau (Intel) indicate that he had run a subjective test with 5 s duration sequences without encountering any problems.[3]

Each 4 s sequence was chosen to have no scene cuts and similar characteristics throughout. That is, the amount of detail and motion are consistent throughout the clip. Private communications with Vittorio Baroncini (GBTech) indicate that the Motion Picture Experts Group (MPEG) uses these characteristics when selecting source sequences to evaluate video codecs. Basically, the

[3] Philip Corriveau performed a subjective video quality test with 5 sec video sequences, while working for the Canadian Research Centre (CRC). The results were published, but we were not able to locate this older document.

goal of each 4 s sequence was to reflect the quality of a 1 or 2 frame sequence. The quality of one or two frames cannot be perceptually perceived or rated, but their quality ratings are nonetheless of great interest to codec developers. Thus, each 4 s sequence is intended to portray a consistent amount of motion throughout, to minimize the quality impact of temporal changes. There are a very small number of exceptions to this rule (i.e., clips where the type of motion or amount of detail changes over the 4 s sequence). These sequences were inserted to provide outliers for identifying metrics that might become unstable in response to such content.

Prior to editing interlaced content, this footage was deinterlaced and format converted using the TMPGEnc© software. The videos were deinterlaced with high precision interpolation, converted to 720p 24fps, and then encoded with H.264 High Profile, CBR, 40 Mbps.

Dataset **its4s** contains the seven HRCs identified in Table 1.

Table 1. HRCs

HRC	**Video Processing Chain**
original (aka SRC)	1) The video was converted to 720p 24fps 2) The video was encoded with H.264 High Profile, VBR, 2-pass at 20 Mbps, to ensure correct playback during the test
SRCpls	1) The video was converted to 720p 60fps 2) The video was encoded with H.264 High Profile, VBR, 2-pass at 20 Mbps, to ensure correct playback during the test
2340K	1) The video was converted to 720p 24fps (1280×720) 2) Video was encoded with H.264 High Profile, VBR, 2-pass at 2.340 Mbps,
1732K	1) The video was converted to 720p 24fps (1280×720) 2) The video was down-sampled to (1024×576) 3) Video was encoded with H.264 High Profile, VBR, 2-pass at 1.732 Mbps
1256	1) The video was converted to 720p 24fps (1280×720) 2) The video was down-sampled to (824×464) 3) Video was encoded with H.264 High Profile, VBR, 2-pass at 1.256 Mbps
0951K	1) The video was converted to 720p 24fps (1280×720) 2) The video was down-sampled to (696×392) 3) Video was encoded with H.264 High Profile, VBR, 2-pass at 0.951 Mbps
0512K	1) The video was converted to 720p 24fps (1280×720) 2) The video was down-sampled to (512×288) 3) Video was encoded with H.264 High Profile, VBR, 2-pass at 0.512 Mbps

SRCpls was included to give limited insight into the drop in quality associated with 24fps content.

The range of encoding bitrates was chosen after considering the current practices of several adaptive streaming providers (e.g., ESPN, Sky Broadcasting, Adobe, YouTube, Georgia Tech, FFXIV, Lighterra), through a mixture of private communications and a review of publicly

available information. The encoding bitrates in Table 1 span the range of bitrates that industry considers (roughly speaking) to be appropriate for 720p streaming.

Notice the linear relationship between encoding bitrate and encoding resolution. The linear relationship between bitrate and resolution is given in (1)-(32), where *bitrate* is the bitrate in Kbps, *vert* is the number of lines vertically, and *horiz* is the number of pixels horizontally. This formula was chosen to approximate the adaptive streaming bitrate ladders available publicly. The bitrates were spread somewhat evenly between 2.34 and 0.512 Mbps, with the constraint that *horiz* and *vert* must both be divisible by eight.

$$pixels = (0.31516 \times bitrate + 222.35)^2 \tag{1}$$

$$vert = \sqrt{\frac{9}{16} \times pixels} \tag{2}$$

$$horiz = \frac{16}{9} \times vert \tag{3}$$

This simplified relationship between bitrate and resolution is unrealistic. Actual adaptive streaming ladders include multiple aspect ratios, match a single resolution with multiple bitrates, and emphasize popular resolutions, like 720p (1280 × 720), widescreen 480p (848 × 480) and VGA (640 × 480). The bitrate ladder in Table 1 eliminates these confounding factors. By consequence, this dataset primarily focuses on the quality response of the original videos.

There are two negative consequences associated with this simplified bitrate/resolution ladder. First, the **its4s** dataset is inappropriate for training any metric that may correlate to video resolution. All video sequences are up-scaled to 720p (1280 × 720) during decoding so, for example, a blurring metric will have a strong response to the encoding resolution. Second, the **its4s** dataset cannot be used to analyze the relationship between resolution, bitrate, aspect ratio, and quality.

Table 2 describes the sessions and the type of content in each session. Table 2 refers to tables in Appendix B for a full description of that session's processed video sequences (PVSs), including footage attribution, licensing terms, editing errors, and file naming convention. The footage was divided into sessions toward the end of the video editing process. Until then, it was not obvious how much of the footage available to ITS fit with the scene selection criteria described in Section 3. Each session has a theme, except for session "chance" which contains the miscellaneous clips remaining. There were two reasons for organizing sessions by theme. The first goal was to relieve boredom. The second goal was to provide context or expectations around the subject matter. The work of Lucjan Janowski and Margaret Pinson (unpublished) raised concerns that subjects might have difficulty rating video sequences that depict unique and unexpected topics.

Table 2. Session Descriptions

Code	Session	Description	RSRC Descriptions
B	Broadcast	Production quality and content typical for broadcast television, including simulated news and movies.	Table B-2
C	Chance	Miscellaneous footage	Table B-3
E	Everglades	Scenes of the Florida Everglades	Table B-4
M	Music & Mexico	Scenes from a music video, a dance sequence, and Mexico.	Table B-5
N	Nature	Various nature scenes (e.g., mountains, cattle drive, Canadian geese, elephants, sea lions and boats)	Table B-6
O	Ocean	Various ocean scenes: waves on the beach and underwater	Table B-7
P	Public Safety	Crowd scenes, mock prison riots, and simulated emergency telemedicine footage	Table B-8
S	Sports	Various sports scenes (e.g., horse race, soccer, skiing, martial arts, skateboarding, boxing, hot air balloon)	Table B-9
T	Training	Various space and technology footage from NASA	Table B-10

Table 3 shows the number of sequences associated with each HRC within each of the eight sessions. Column “Code” is the single letter session code used in the video file naming convention. The original video sequences are divided into two columns: “original good” and “original fair”. These divide the original videos into good quality or better, and fair quality or worse, respectively. These were the quality judgements of the author when editing the videos.

Table 3. Distribution of HRCs Per Session

Code	Session	Original Good	Original Fair	SRCpls	2340K	1732K	1256K	0951K	0512K
B	Broadcast	29	3	3	13	12	11	16	14
C	Chance	30	0	1	12	10	15	17	15
E	Everglades	33	0	0	12	10	11	19	15
M	Music & Mexico	32	5	7	10	9	10	16	11
N	Nature	29	4	0	10	10	11	14	12
O	Ocean	25	9	3	10	10	13	16	14
P	Public Safety	26	1	3	14	12	14	15	15
S	Sports	25	1	5	13	12	12	19	13
T	Training	2	1	0	2	2	1	2	2

The experiment was divided into eight (8) sessions, each containing approximately100 sequences. Each session implemented the same approximate experiment design with a different type of video content. Part of our experimental goal was to determine whether subjects responded positively to this change in scene content from one session to another. One session contained first responder content, and our goal with that session was to determine if subjects drawn from the populace at large would object to rating this content or if they would show any adverse scoring behaviors. Best practices in the design of subjective video quality experiments is

to avoid all content that may trigger strong emotional responses or biased opinions. The first responder content falls into this category, as it contains firearms and simulated wounds.

Most of the video content was edited from footage gathered by ITS between 2004 and 2015. In some cases, the 4 s clips were edited from raw video material, where an edited version of that sequence has been made available on CDVL (www.cdvl.org). Some footage was made available by other organizations, who allow their footage to be redistributed. Since none of the footage was filmed with this experiment in mind, a varying number of video clips with similar properties was gathered from each content type. Thus, each RSRC is associated with an uncontrolled number of sequences.

The clips within each session were assigned to HRCs using a combination of random chance and constrained balance. If an RSRC contained many clips, the association of sequence to HRC was constrained to ensure one clip was matched with each HRC (or at least to come close to that ideal). When an RSRC contained footage appropriate for the SRCpls impairment (i.e., filmed at 59.94 frames or fields per second), then a clip with very similar subject matter was assigned to the original HRC as well, to allow for direct quality and material comparisons. Likewise, a few pairs of sequences with similar characteristics were assigned to original (aka SRC) and 2340K, to allow comparisons between the original video and the bitrate that industry considers "transparent" for consumer applications. Ugly SRC (i.e., with fair quality or less) were always assigned to the original HRC, with the goal that roughly 5 of the 100 PVSs in each session should be "ugly SRC." The actual number are lower, because some content types contained no such sequences.

Other than the above constraints, clips were randomly assigned to HRCs with the intention that each session should have approximately the distribution identified in Table 4. Note that the ugly SRC are simply labeled "SRC" or "original" in the actual experiment's data files.

Table 4. Intended Distribution of HRCs per Session

Name	**100 clips**
SRCpls	5
orginal	35
Ugly SRC	5
2340K	10
1732K	10
1256K	10
0951K	15
0512K	10

Essentially, we selected the application consumers loosely and ambiguously call "high definition video on a mobile device." Consumers typically use "high definition" to describe their quality expectations, unaware of the video resolution connotation. The application determined the video format (720p 24fps) and the range of bitrates. Bitrates from 0.9 to 2.34 Mbps are appropriate for a contemporary 720p broadcast service, and minimum bitrate (512 Kbps) characterizes low bitrate outliers. The impairment level 0951K was assigned slightly more clips, because industry

experts indicated that 720p would rarely be streamed at bitrates below 900 Kbps. The original HRC is associated with many more video clips than the other HRCs, because a primary purpose of this experiment is to understand the diverse quality responses produced by many different content types.

Between random chance and limitations on subject material, each of the 8 sessions contains a slightly different distribution of clips among HRCs. Some types of content contained no ugly SRCs; and other types of content contained no material suitable for creating the SRCpls impairment.

The experiment used a modified version of the ACR rating scale. In addition to the standard five level scale (excellent, good, fair, poor, bad), two new response levels were available: “human error” and “computer error.”

Subjects were instructed to select “human error” if they could not rate the clip. For example, they were distracted and did not pay attention to the video. The motivation was twofold. While participating in prior subjective tests with 8 to 12 s sequences, we have experienced such lapses in attention, so we expect other subjects to be likewise distracted. The standard ACR method forces the subject to choose, so distracted ratings add noise to the data. Thus, our first motivation was to investigate the impact of a “human error” option on the rating scale. Additionally, we were concerned that some subjects might find the fast pace of the experiment uncomfortable, or that some sequences would be difficult for all subjects to rate. The “human error” level provided a mechanism for subjects to help us identify whether the 4 s sequence duration was sufficient (i.e., by subject and by sequence).

Subjects were instructed to select “computer error” if our playback system had a problem. For example, they saw a flash of a different video at the beginning or end of the video. Subjects were told that they should not encounter any computer errors. The “computer error” served two purposes. The first was intended to identify any editing errors that our quality control process failed to detect. The second was to identify unknown or sporadic flaws in our video playback system.

Subjective data was collected at both ITS and AGH University, in Krakow Poland.

At ITS, the experiment was run on laptops with 720p resolution. The video playback, randomization and data collection were supported by the WEST software [9]. The experiment began with the training session, which contains 12 sequences. The ITS subject pool contained six ITS employees, engineers who took the laptop to their office and took the test at their leisure. Each office provided a quiet environment with a mixture of natural and artificial lighting. The first three subjects were allowed to self-select session ordering according to their interest in the subject matter. All three chose sessions in alphabetical order, so the remaining subjects were given cards with the names of the eight sessions and instructed to use those cards to randomly choose session ordering.

The other 21 ITS subjects were provided by a temporary hiring agency. These subjects were run through the experiment two at a time in a quiet room with natural lighting. Appendix A contains the instructions that ITS read to the subjects. The experimenter was present at all times, on the

opposite side of the table. All ITS subjects self-reported their vision on the ACR scale. Overall, the ITS subject pool contains data from 12 males, 14 females, and one person whose demographic data was lost.[4]

The whole experiment was self-paced, both within each session and between sessions. Subjects were encouraged to take a break at any time, instead of enforcing breaks between sessions. Subjects often chose not to pause between sessions, and all subjects occasionally paused in the middle of a session. Subjects took an average of 15 min to rate each session, and the entire test duration ranged from 2 to 2.5 hours.

AGH University collected data using 24 engineering students who were learning about subjective testing. The subjects took the experiment in a laboratory with 10 workstations. Five of these students were female, 19 were male and their ages ranged from 20 to 25 years. The students rated two of the eight sessions: Everglades and Sports.

Not noted in the above tally are two ITS subjects from the temporary hiring agency who were rejected during subject screening. The data was screened using the Pearson correlation method from Annex A.1 of ITU-T Rec. P.913. We used a rejection threshold of 0.4, which rejected a female with bad vision (correlation 0.17) and a male with good vision (correlation 0.10). Discussions with the male after the experiment indicated that he did not understand the instructions. All other subject correlations were 0.62 or above. The low rejection threshold reflects our desire to err on the side of retaining subjects, due to several unusual design choices.

During the instructions, ITS warned subjects that the Public Safety session would contain simulated wounds and guns firing simulated teargas. Subjects were encouraged to skip or discontinue this session if the content bothered them.

[4] This subject was an ITS employees, whose age is within the range of ages for other ITS subjects.

4. MOS ANALYSIS

This section analyzes the **its4s** dataset MOSs. The goal was to analyze whether the experiment design caused problems. Put briefly, we found no problems with the experiment design that were not known before collecting data from subjects (e.g., the unrepeated scenes, RSRC, 4 s clip duration, and first responder content did not produce any unexpected problems). The known problems are a consequence of the design trade-offs discussed in Section 3.

The raw ratings from ITS and AGH are pooled (i.e., without scaling). Janowski and Pinson [10] provide a formal analysis supporting this procedure, which is commonly used when a single experiment is split among two or more labs.

4.1 AGH vs ITS

Figure 1 shows the relationship between AGH MOSs and ITS MOSs for the two sessions rated by AGH (Everglades and Sports). The lab-to-lab correlation is 0.94, which is within the range of lab-to-lab correlations commonly seen in prior experiments. The lower average MOS of the AGH data may reflect the younger age of AGH subjects (i.e., superior eyesight).

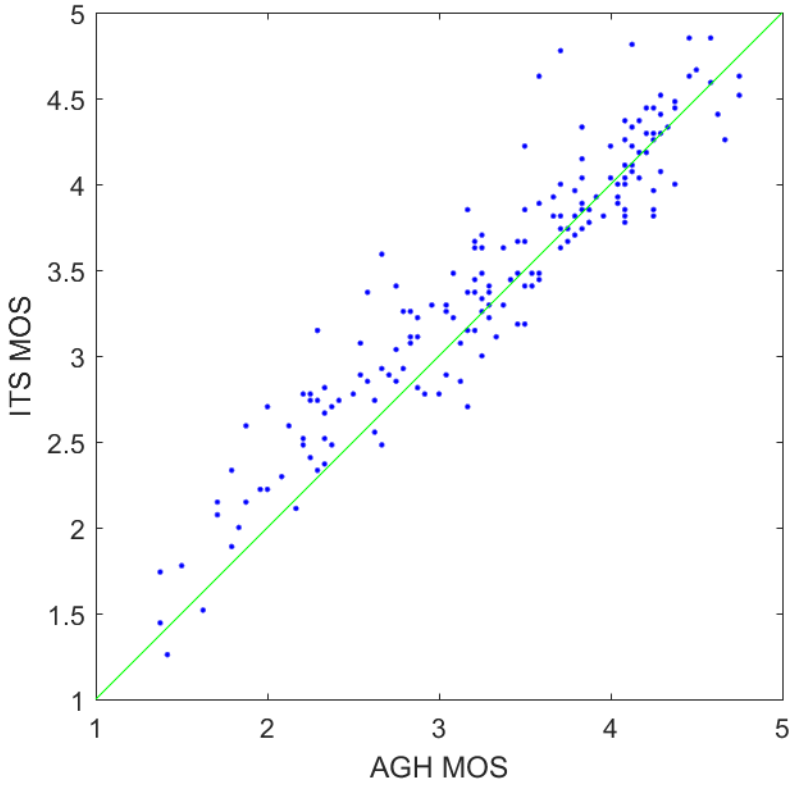

Figure 1. Scatter plot relating MOSs from AGH University (x-axis) and ITS (y-axis) indicates a good lab-to-lab correlation.

4.2 Human Error Rating Level

Recall that the 27 ITS subjects were asked to select "human error" if they could not rate the clip. Table 5 shows how often subjects used this rating option—33% of subjects never used the "human error" rating level. The subject who selected "human error" most often only used this option for 2.3% of the data. No sequence was associated with more than one "human error" rating. AGH did not include the "human error" and "computer error" options.

Overall, we see no evidence that subjects had problems rating 4 s sequences. The "human error" response level appeared to add value to the rating scale with minimal impact.

Table 5. "Human Error" Rating Level.

Times Used	**0**	**1**	**2**	**3 to 5**	**6 to 10**	**19**
# Subjects	9	4	3	6	4	1

4.3 Computer Error

The "computer error" rating level was never selected, despite 2.5% of video sequences unintentionally having a few frames from another sequence at the beginning (see Table B-11). This could be an artifact of human perception (e.g., a reduced ability to observe immediately after the scene cut denoting the beginning of the sequence playback), but it is also possible that the playback system always skipped the first few frames.

4.4 ITS Subject Feedback

After the experiment concluded, ITS asked subjects for feedback on the experiment design. Most subjects expressed interest in the content and how ITS obtains footage for this research. For examples:

- "Is Naomi a real person?"
- "I pieced together the plot for Tears of Steel."
- "They did a good job on the simulated wounds."
- "Where did you get the telemedicine footage?"

Subjects provided positive feedback on the topic changes between sessions, the public safety content, 4 s sequence duration, and the length of the entire experiment.

Two subjects complained that there were too many video clips for a given content type. These content types were bees (17 sequences) and Canadian geese (44 sequences). Note that some of the other content types did not trigger such complaints, despite similar numbers. The most notable is the Everglades session, which was drawn entirely from one initial footage source.

Other examples are crowd sequences, telemedicine, and mock prison riots, each of which comprise around one third of the Public Safety session.

None of our subjects provided negative feedback on the experiment duration. Initially, we considered splitting the eight sessions among two pools of subjects, to limit each subject's participation to under 1.5 hours. Feedback from preliminary testing had indicated that the content variety, 4 s sequences, and fast pace reduced fatigue and would allow subjects to rate all eight sessions in under three hours. The preliminary testing data was discarded.

4.5 MOS Distribution

Figure 2 shows the distribution of MOSs for each HRC, pooling all data from all sessions. Table 6 shows HRC MOSs, computed for the **its4s** dataset and for each individual session. Session HRC MOSs are omitted for SRCpls, because there are insufficient samples.

Of the 257 original video sequences, only 141 (56%) received a quality rating of good or better (MOS ≥ 4). During editing and scene selection, our goal was that 35 of 40 original videos (87.5%) would have a quality of good or better. Based on this selection criterion and knowledge from prior experiments, we predicted that the original HRC MOS (average MOS across all original sequences) would be around 4.25.

Instead, the original HRC MOS was 4.01. That is, the original HRC MOS was shifted downward from our expectations during scene selection. We have observed similar shifts in [1] and other VQEG validation test datasets (i.e., experts judged an original video to have quality good or better, but the MOS was < 4.0). The **its4s** dataset's "original HRC" contains 24fps content, which may explain some of the downward shift (e.g., we ignored 24fps artifacts as a consequence of this design choice, while subjects noticed and disliked the jerky motion).

The SRCpls HRC MOS was 4.34, and 19 of the 23 SRCpls sequences (83%) had a quality of good or better. The increase in frame rate from 24 fps to 60 fps causes this increase of 0.33 MOS. The SRCpls distribution is very similar to the distribution of original sequences in the six VQEG HD datasets [1]. We can see this by comparing histograms of the VQEG HD original sequences with the **its4s** dataset's original and SRCpls sequences, as shown in Figure 3. The similar distribution of the VQEG HD dataset SRCs addresses our concerns around SRCpls (i.e., small sample size, still or nearly still scenes are underrepresented).

Table 6. HRC MOSs by Session.

Session	Original	SRCpls	2340K	1732K	1256K	0951K	0512K
Broadcast	4.14	—	3.86	3.78	3.54	3.18	2.35
Chance	3.97	—	3.73	3.59	3.14	3.25	2.20
Everglades	3.91	—	3.60	3.54	2.99	2.67	1.91
Music & Mexico	4.09	—	4.00	.376	3.45	3.14	2.47
Nature	4.14	—	4.02	3.70	3.36	2.95	2.22
Ocean	3.90	—	4.19	3.83	3.05	2.98	2.30
Public Safety	3.93	—	3.71	3.49	3.30	2.85	2.33

Session	Original	SRCpls	2340K	1732K	1256K	0951K	0512K
Sports	3.95	—	3.77	3.61	3.29	3.04	2.34
Whole Dataset	4.01	4.34	3.85	3.66	3.27	2.99	2.27

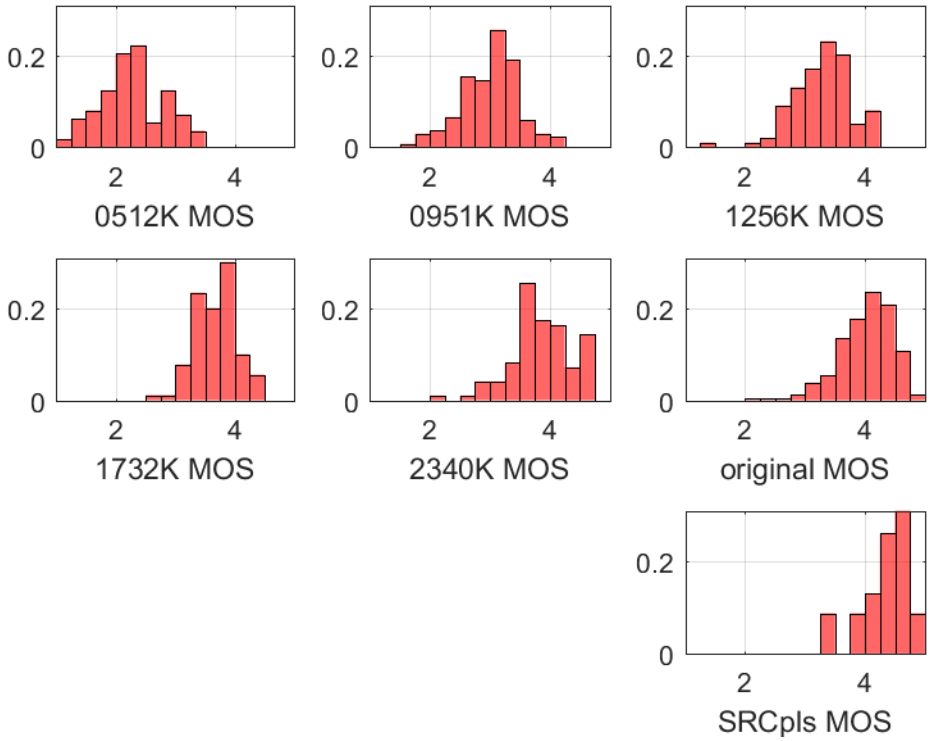

Figure 2. Histograms showing the distribution of MOSs for each HRC.

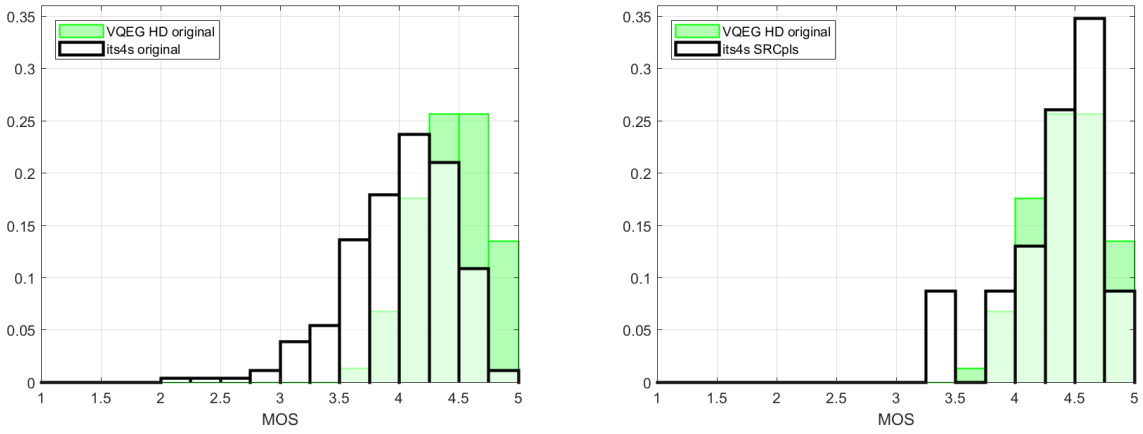

Figure 3. Histograms comparing the distribution of original MOSs the six VQEG HD datasets (green) with the **its4s** dataset's 24fps original videos (left) and the **its4s** dataset's 60 fps SRCpls videos (right). The **its4s** dataset histogram is overlaid with a thick black outline.

4.6 HRC MOS Distribution

This section compares sessions based on conclusions about HRCs. The HRC MOS for a session is the average MOS across all sequences associated with that session and HRC. The SRCpls HRC is omitted from this section's distribution analyses, due to the small sample size.

Figure 4 plots one session's HRC MOS (x-axis) against the HRC MOS averaged over all eight sessions (y-axis). We calculate the average over sessions (instead of pooling individual sequences) to reduce the impact of the differing number of sequences associated with any particular HRC, depending upon the session. The value "Δ" within x-axis label indicates the overall bias (shift) between that session's MOSs and the experiment as a whole. Figure 4 shows overall biases, such that some sessions' sequences receive slightly higher or lower scores. This is unsurprising, given variations among production quality and subject matter.

Table 7 shows Pearson correlation between HRC MOSs, calculated for all session pairs. These correlations are typically very high except when the Ocean session is included in the pair. Notice that the Ocean session contains nearly twice the number of ugly SRCs as any other session (see Table 3). This likely causes the increased scatter we see the high end of the scale of the Ocean session's scatter plot (see Figure 4). Nonetheless, the lowest of these session-to-session correlations are within the range of lab-to-lab correlations (i.e., when two or more different labs rate the same video sequences).

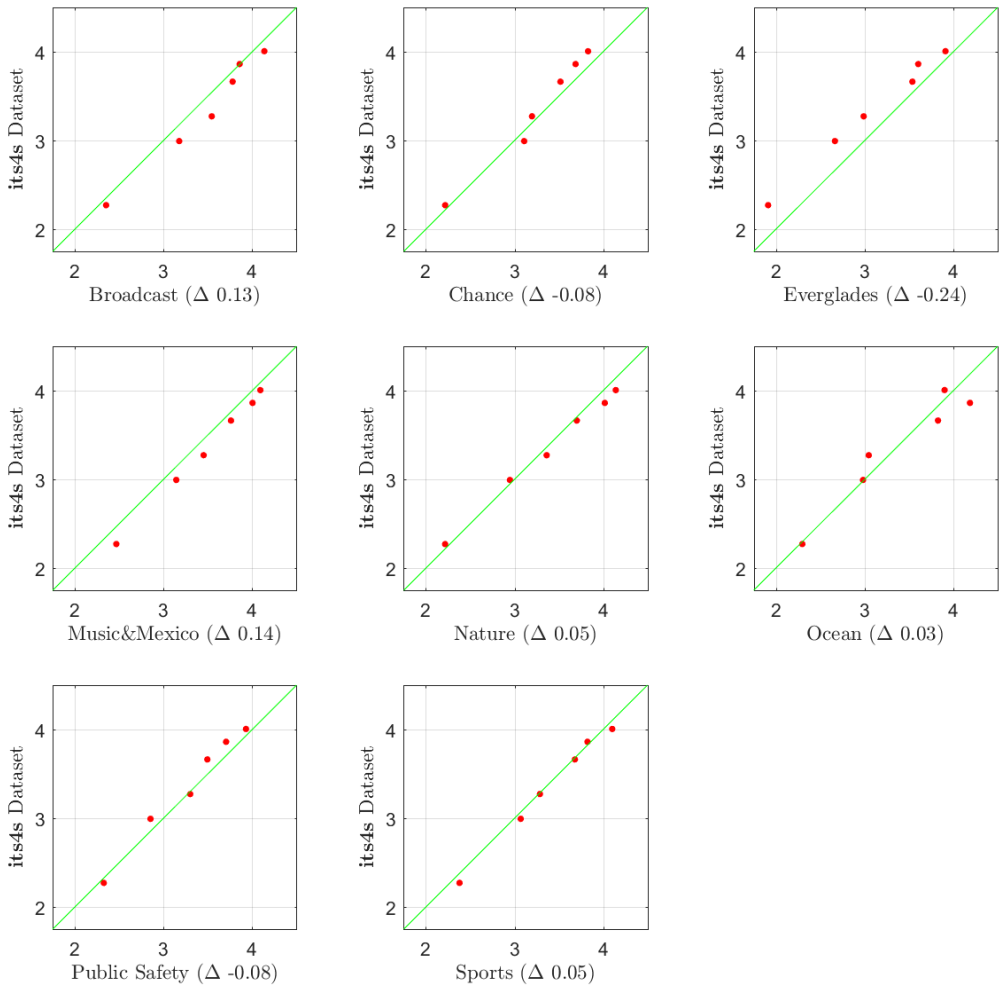

Figure 4. Scatter plots show the relationship between each session's HRC MOSs (x-axis) and HRC MOSs calculated from all session MOSs (y-axis).

Table 7. Correlation Between Each Session's HRC MOSs

	Broadcast	**Chance**	**Everglades**	**Music&Mexico**	**Nature**	**Ocean**	**Public Safety**	**Sports**
Broadcast	1.00	0.99	0.99	0.99	0.99	0.92	0.99	0.99
Chance	0.99	1.00	0.98	0.99	0.98	0.95	0.97	0.99
Everglades	0.99	0.98	1.00	0.99	0.99	0.96	0.99	1.00
Music&Mexico	0.99	0.99	0.99	1.00	1.00	0.96	0.99	0.99
Nature	0.99	0.98	0.99	1.00	1.00	0.96	1.00	0.99
Ocean	0.92	0.95	0.96	0.96	0.96	1.00	0.93	0.95
Public Safety	0.99	0.97	0.99	0.99	1.00	0.93	1.00	0.99
Sports	0.99	0.99	1.00	0.99	0.99	0.95	0.99	1.00

We can also use the Student's *t*-test to evaluate whether the set of MOSs associated with one HRC and session come from the same distribution as the MOSs associated with that HRC from the other seven sessions. This Student's *t*-test examines the distribution of PVS MOSs, not the distribution of subject ratings. Each set of PVS MOSs represents the larger set of MOSs for all possible 4 s original videos, processed as per Table 1. From the point-of-view statistics, the accuracy of the Student's *t*-test does not depend upon the number of subjects.

At the 95% confidence level, the Student's *t*-test indicate a different distribution for 6 of these 48 comparisons (12.5%). These are Everglades 0512K, Everglades 0951K, Broadcast 1256K, Everglades 1256K, Ocean 2340K, and Chance original. Four of these six differences stem from the Everglades session. Notice that the Everglades session scatter plot (see Figure 4) indicates a slight linear gain that is not seen in the other sessions (i.e., the HRC MOSs on the x-axis span a wider range than the HRC MOSs on the y-axis). Probably the Everglades content contains some unique factor or predominance of content. One possibility is moving water, but that characteristic dominates both the Ocean and Everglades sessions. A more likely characteristic is unusual camera movement: many of the Everglades sequences were filmed from a moving boat.

These HRC MOS analyses establish that the **its4s** dataset is sensitive and robust. The extremely high correlations, tight scatter plots, and Student's *t*-test conclude that all eight sessions characterized the HRCs very similarly. This is strong evidence that the unrepeated scene design is well suited for comparing HRCs. The only caveat is that the experiment design must match a sufficiently large number of scenes for each HRC.

4.7 Public Safety Session

No subject skipped or discontinued the public safety session. The analyses in Section 4.6 indicate that the Public Safety session tracks the response of the other seven sessions. Discussions with subjects after the experiment indicated that some subjects believed the telemedicine footage was real, despite the instructions explicitly stating that these were simulated wounds.

The conventional wisdom to avoid polarizing content might be set aside when there is a genuine research need. Caution would be still advisable when choosing and instructing subjects, particularly when showing content that is more disturbing than the footage appearing in the **its4s** dataset.

5. NR METRIC ANALYSIS

The ultimate goal of the **its4s** experiment design was to provide means to improve NR metrics. Thus, it is appropriate to analyze the dataset from this perspective. Since **its4s** focused on qualities of original content, we will present a metric that likewise focuses on the original video—as opposed to impairments associated with transmission or storage compromises. The purpose of this section is to demonstrate the type of NR metric that the **its4s** dataset was designed to train.

From a theoretical standpoint, we would like our NR parameter to have the following characteristics:

- The parameter tracks some aspect of quality that varies across the original content. That is, we want the NR parameter to be theoretically plausible, given our understanding of aesthetics, composition, etc.

- The accuracy of the parameter does not degrade as bitrate drops. That is, the parameter measures one or more intrinsic properties of the video that people always care about, regardless of what other impairments are introduced.

- The parameter is not influenced by coding impairments. For example, the same values are calculated for both the original video and the 512 Kbps compressed video. This is an unusual requirement, as most FR, RR, and NR metrics focus primarily or completely on coding impairments. However, an important goal of the **its4s** dataset is to accurately predict the quality of video sequences that do not contain coding artifacts.

- A scatter plot between MOS and the NR parameter should cover an upper or lower triangle. That is, as the parameter value approaches zero, the MOSs will span the full range from excellent to bad; but as the parameter value increases, the range of MOSs will diminish and approach one end of the scale (excellent or bad). In other words, when the parameter is zero it provides no information but as the parameter increases it sets increasingly more stringent lower or upper bounds on MOS. As we build up a family of such parameters, each tracking different perceptually relevant attributes, we should be able to place tighter and tighter limits on estimated MOS.

Notice that the fourth criterion is very different from FR or RR metric development, where we seek parameters that scatter evenly around the regression line. Such parameters always track all types of quality impairments, at least in theory. For NR metric development, we will probably need many different parameters, most of which will be irrelevant at any given time (e.g., lens flare, lens distortion, panorama stitching errors, camera focusing errors). By consequence, we expect that most of these NR parameters will have a fairly low correlation to MOS.

5.1 Munsell Color Space

Our NR parameter builds upon the study of color presented in [11]. In that paper, we analyzed color from multiple perspectives and questioned whether the color spaces popularly used for

video coding and objective video quality metrics were well suited for these purposes. The reader is guided to Section 4 of [11] to understand those theoretical underpinnings.

The Munsell color space was designed by an art instructor, Albert H. Munsell, who built upon prior work by, for example, physicist Herman von Helmholtz (see [12]). The Munsell color space [13] describes colors using hue, value, and chroma. Hue indicates color and is quantized into 40 discrete values around a circle. Hues are coded by a two level scheme of letters then numbers. Value indicates lightness or darkness and ranges from 0 to 10. Chroma indicates the saturation or brilliance of a color and ranges from 0 to 26. The Munsell color space is asymmetrical, as contrasted to RGB or YCbCr, which are assumed to span a tidy cube with limits corresponding to a power of two. The Munsell color space prioritizes human perception of color over mathematical convenience.

Thus, the transformation from the YCbCr (or RGB) color space to the Munsell color space involves a lookup table. We used the table provided by Rochester Institute of Technology (RIT) [14], which is based on the report of Newhall, Nickerson, and Judd [15] in 1943. Judd and Nickerson issued an updated report in 1967 [16] that supposedly offered a huge improvement in perceptual uniformity. However, those improved values were only available as a scan of the National Bureau of Standards (NBS) report. We chose to wait to pursue that until we need the improved uniformity. RIT would appreciate hearing back from anyone who uses the 1967 values.

The RIT spreadsheet omits neutral values. We added RGB values for neutral colors N1 through N9, as defined by the interactive tool provided by Andrew Werth [17]. Munsell white (N10) and Munsell black (N0) cannot be represented in the RGB color space. Nonetheless, we added approximate RGB values for N0 at computer black (R=0, G=0, B=0) and ITU-R Rec. BT.709 [18] reference black (R=16, G=16, B=16). Likewise, we added approximate values for N10 (white) at both computer white (R=255, G=255, B=255) and ITU-R Rec. BT.709 reference white (R=235, G=235, B=235).

5.2 Munsell Red NR Parameter

We will evaluate the **its4s** dataset with a simple metric that measures the fraction of red pixels. The importance of the color red can be understood by reading the work of linguists, visual psychologists, and anthropologists to understand human perception of color (see [11]). More colloquially, there is the long-time belief that red has a huge impact in still photography.

The *Munsell Red* NR parameter was computed as follows:

1) Convert YCbCr pixels to Munsell using a nearest neighbor search.

2) For each frame, compute the fraction of pixels with chroma > 2 and hues between 2.5YR and 2.5P (inclusive). This region of the color space is described in more visual terms below.

3) Compute mean over all frames.

4) Apply the square root.

The thresholds in step 2 were selected by examining the Munsell naming convention and visually examining Munsell colors on a monitor. Roughly speaking, this hue range includes purple, red-purple, red, and some shades of orange, while the chroma range excludes greyed colors. This range of hues is based on a simplified color naming system that includes seven colors (i.e., black, white, grey, yellow, red, green, and blue). This simplified color naming system is justified by prior work on the order in which color terms enter languages [11]. The thresholds in step 2 were not optimized for the **its4s** dataset and are somewhat arbitrary (e.g., different people would shift the red/blue boundary from 2.5P to another hue). Thus, *Munsell Red* is the mean fraction of pixels that lie in a specific region of the Munsell color space. *Munsell Red* ranges from 0 (no pixels are in this region) to 1.0 (all pixels are in this region).

Only the "mean" temporal collapsing function was considered for step 3. This is due to the philosophical consideration that we want to be able to apply the NR parameter to a single frame of video, to meet the needs of product developers. Recall that the **its4s** dataset was designed to minimize temporal variations.

The square root was indicated by plotting the data (NR metric vs MOS).

To analyze the *Munsell Red* NR parameter, we will use reference videos. By this we mean the video as it appeared after step 1 in Table 1 (conversion to 720p) but before downsampling and compression. For the original and SRCpls HRCs, there is no functional difference between the reference video and the video that appeared in the **its4s** dataset. For the other HRCs, these reference videos were not viewed or rated.

First, we calculated *Munsell Red* on both the reference video and the processed video sequences (PVS), by which we mean the sequences that appear in the **its4s** dataset. The correlation between these is 0.9977. This demonstrates that *Munsell Red* is not influenced by coding impairments.

Second, we calculated the performance of *Munsell Red* on each HRC and the full dataset. Tables 8 and 9 analyzed the metric performance using Pearson correlation and root mean square error (RMSE), respectively. For the RMSE analysis, the metric data were fitted to the MOSs with a simple linear fit. In each table, the row "PVS" contains analyses based on the metric calculated on the PVSs that appear in the **its4s** dataset; and row "Reference" contains analyses based on the metric calculated on the reference videos. Pearson correlation is ill-suited to this analysis, because correlation drops as the range of data narrows (see Section III.D of [19]). This table is included for completeness. Table 9 indicates that the performance of the parameter does not degrade as bitrate drops. Figure 5 plots *Munsell Red* vs MOS for the entire **its4s** dataset and Figure 6 plots *Munsell Red* vs MOS for each HRC. This plot shows that the goal of an upper triangular distribution has been achieved. Small values of *Munsell Red* provide no information about MOS, but as *Munsell Red* increases it provides an approximate lower bound on MOS. For example, very few MOSs are less than (4.0 × *Munsell Red*).

Finally, we will analyze the performance of *Munsell Red* on the consumer content resolution and image quality dataset (CCRIQ) [20]. This dataset and subject ratings are available on the CDVL (www.cdvl.org). The CCRIQ dataset contains 18 scenes, each photographed with 23 different cameras. The CCRIQ images were scaled to the resolution of the viewing monitor before calculating *Munsell Red*. Since the images were rated separately on 4K and HD (1920 × 1080)

monitors, this doubles the amount of data. The Pearson correlation between *Munsell Red* and the CCRIQ MOSs is -0.07. When we examine individual cameras (per resolution), correlation values range from 0.37 to -0.62. Using RMSE as a metric, the overall performance on CCRIQ is 1.01, with the RMSE ranging from 0.25 to 1.17 for individual cameras, though these values optimistically ignore the negative correlation problem. Figure 7 plots *Munsell Red* vs MOS for the CCRIQ dataset.

Basically, *Munsell Red* performs very poorly on the CCRIQ dataset. This is not intrinsically discouraging, since our goal was to demonstrate an NR parameter that was theoretically plausible but did not respond well to prior datasets. The CCRIQ dataset has 2.2% of the source material variety of the **its4s** dataset, so small biases in content choice would be highly problematic. Another difference is that most of the **its4s** dataset was produced by profession videographers (see Table B-1), while the CCRIQ images were photographed by amateurs. NR metrics may, at least in the near future, require a flag that helps the metric understand the differential impact of amateur and professional camera operators.

Table 8. Pearson Correlation Analysis of *Munsell Red*

Video	**Original**	**SRCpls**	**2340K**	**1732K**	**1256K**	**0951K**	**0512K**	**Full Dataset**
PVS	0.17	0.12	0.15	0.28	0.20	0.32	0.39	0.16
Reference	0.17	0.13	0.15	0.25	0.18	0.31	0.38	0.15

Table 9. RMSE Analysis of *Munsell Red*

Video	**Original**	**SRCpls**	**2340K**	**1732K**	**1256K**	**0951K**	**0512K**	**Full Dataset**
PVS	0.44	0.40	0.48	0.33	0.47	0.46	0.48	0.77
Reference	0.44	0.40	0.48	0.34	0.47	0.47	0.49	0.77

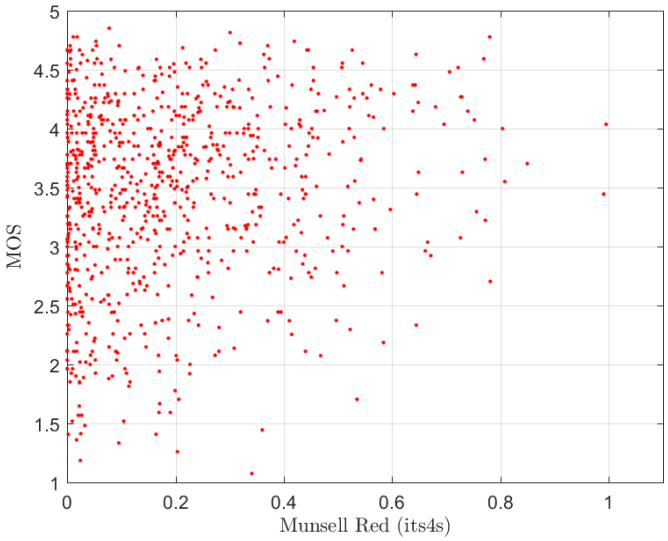

Figure 5. Performance of the NR metric *Munsell Red* on the **its4s** dataset.

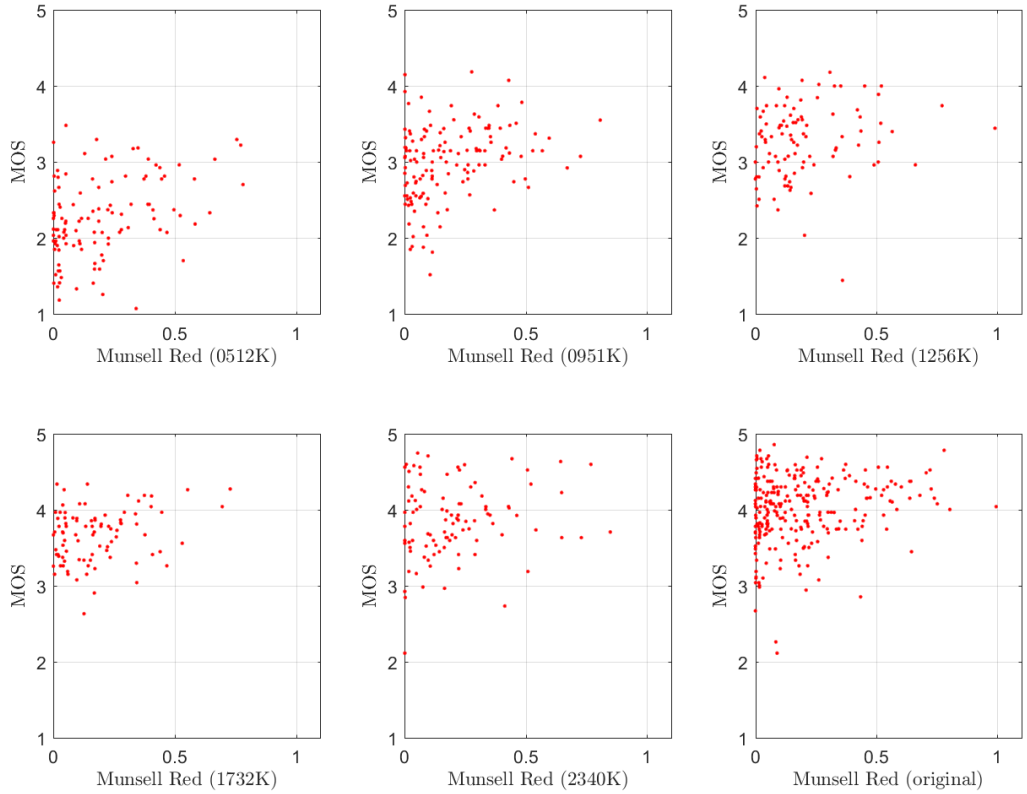

Figure 6. Performance of the NR metric *Munsell Red* on each HRC within the **its4s** dataset.

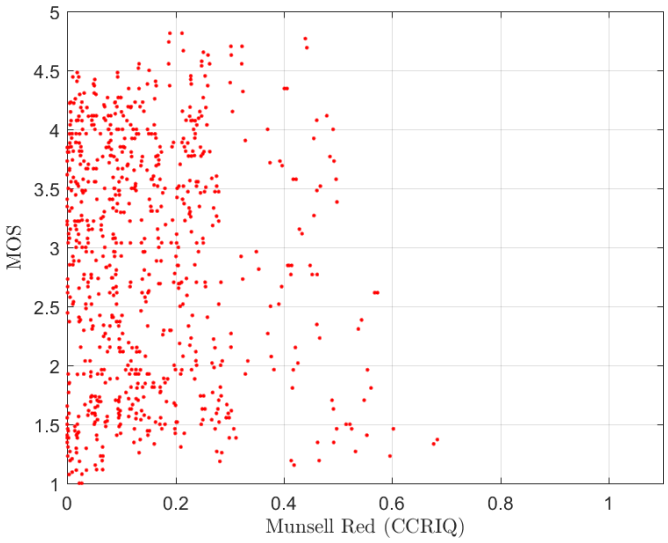

Figure 7. Performance of the NR metric *Munsell Red* on the **CCRIQ** dataset.

6. CONCLUSIONS

The **its4s** dataset offers an alternative experiment design that is intended specifically to inspire innovative NR metric development. The experiment design emphasizes a large variety of unrepeated scenes and short duration video sequences (4 s). The **its4s** dataset includes poor quality footage, which is intentionally omitted from most subjective video quality experiments. The NR parameter *Munsell Red* demonstrates the type of NR metric development that might be possible using this dataset and others designed similarly. The **its4s** dataset indicates that reducing a video's frame rate from 60fps to 24 fps causes a decrease of ≈0.33 MOS.

The **its4s** dataset is available on the Consumer Digital Video Library (CDVL, www.cdvl.org) for research and development purposes. CDVL provides the compressed video files (as viewed by subjects) and the raw subjective ratings.

7. REFERENCES

[1] Video Quality Experts Group (VQEG) report on the validation of video quality models for high definition video content, Video Quality Experts Group (VQEG), June 30, 2010. <https://www.its.bldrdoc.gov/vqeg/projects/hdtv/hdtv.aspx>

[2] Margaret H. Pinson; Lucjan Janowski; Zdzislaw Papir, "Video Quality Assessment: Subjective testing of entertainment scenes," *IEEE Signal Processing Magazine*, vol. 32, no. 1, pp. 101-114, January 2015. doi: 10.1109/MSP.2013.2292535 <https://www.its.bldrdoc.gov/publications/2821.aspx>

[3] Margaret H. Pinson; Lucjan Janowski, "AGH/NTIA: A Video Quality Subjective Test with Repeated Sequences," NTIA Technical Memo TM-14-505, June 2014. <https://www.its.bldrdoc.gov/publications/2758.aspx>

[4] Stephen D. Voran, "An iterated nested least-squares algorithm for fitting multiple data sets," NTIA Technical Memo TM-03-397, October 2002. <https://www.its.bldrdoc.gov/publications/2428.aspx> Software available at <https://www.its.bldrdoc.gov/resources/video-quality-research/inlsa.aspx>

[5] Margaret H. Pinson; Stephen Wolf, "Techniques for Evaluating Objective Video Quality Models Using Overlapping Subjective Data Sets," NTIA Technical Report TR-09-457, November 2008. <https://www.its.bldrdoc.gov/publications/2494.aspx>

[6] Margaret H. Pinson; Stephen Wolf, "A new standardized method for objectively measuring video quality," *IEEE Transactions on Broadcasting*, vol.50, no.3, pp. 312-322, Sept. 2004. doi: 10.1109/TBC.2004.834028 <https://www.its.bldrdoc.gov/publications/2576.aspx>

[7] Margaret H. Pinson and Stephen Wolf , "A new standardized method for objectively measuring video quality," *IEEE Transactions on Broadcasting*, vol.50, no.3, pp. 312-322, Sept. 2004. doi: 10.1109/TBC.2004.834028 <https://www.its.bldrdoc.gov/publications/2576.aspx>

[8] Margaret H. Pinson, "Technology Gaps in First Responder Cameras," NTIA Technical Memo TM-17-524, May 2017. <https://www.its.bldrdoc.gov/publications/3171.aspx>

[9] Andrew A. Catellier; Luke Connors, "Web-Enabled Subjective Test (WEST) Research Tools Manual," NTIA Handbook HB-14-501, January 2014. <https://www.its.bldrdoc.gov/publications/2748.aspx> Software available at <https://www.its.bldrdoc.gov/resources/video-quality-research/web-enabled-subjective-test-west.aspx>

[10] Lucjan Janowski and Margaret H. Pinson, "The accuracy of subjects in a quality experiment: a theoretical subject model," *IEEE Transactions on Multimedia*, vol. 17, no. 12, December 2015, pp 2210-2224. <https://www.its.bldrdoc.gov/publications/2814.aspx>

[11] Margaret H. Pinson, "A missing factor in objective video quality models: A study of color," *Ninth International Workshop on Video Processing and Quality Metrics for*

Consumer Electronics - VPQM 2015, Chandler, AZ, February 5-6, 2015. <https://www.its.bldrdoc.gov/publications/2786.aspx>

[12] Rolf G. Kuehni and Andreas Schwarz, "Color Ordered, *A Survey Of Color Order Systems From Antiquity To The Present*, Oxford University Press, Inc., 2008.

[13] "Color Theory and History," Official Site of Munsell Color, © 2017, X-Rite, Inc., <http://munsell.com/color-blog/category/color-theory-history/>, accessed on Nov. 14, 2017.

[14] Val Hemink, "real.dat: by the book," Munsell Renotation Data, <https://www.rit.edu/cos/colorscience/rc_munsell_renotation.php>, accessed on Mar. 14, 2017.

[15] Sidney M. Newhall, Dorothy Nickerson, and Deane B. Judd, "Final Report of the O.S.A. Subcommittee on the Spacing of the Munsell Colors," *Journal of the Optical Society of America (JOSA)*, vol. 33, issue 7, 1943, pp. 385-418.

[16] Deane B. Judd and Dorothy Nickerson, "One Set of Munsell Re-renotation," National Bureau of Standards (NBS) Report 192693, December 26, 1967. Available: <http://www.rit-mcsl.org/MunsellRenotation/MunsellRe-renotations.pdf>

[17] Andrew Werth, "Virtual Munsell Color Wheel," <http://www.andrewwerth.com/aboutmunsell/>, accessed Mar. 14, 2017.

[18] Recommendation ITU-R BT.709 (06/2015), *Parameter values for the HDTV standards for production and international programme exchange*, International Telecommunication Union, Geneva, Switzerland.

[19] Margaret H. Pinson; Lucjan Janowski; Romuald Pépion; Quan Huynh-Thu; Christian Schmidmer; Philip J. Corriveau; Audrey Younkin; Patrick Le Callet; Marcus Barkowsky; William Ingram, "The Influence of Subjects and Environment on Audiovisual Subjective Tests: An International Study," *IEEE Journal of Selected Topics in Signal Processing*, Vol. 6, No. 6, October 2012, pp. 640–651. <https://www.its.bldrdoc.gov/publications/2682.aspx>

[20] Michele A. Saad; Margaret H. Pinson; David G Nicholas; Niels Van Kets; Glenn Van Wallendael; Ralston Da Silva; Philip J. Corriveau; Ramesh V Jaladi, "Impact of Camera Pixel Count and Monitor Resolution Perceptual Image Quality," *Colour and Visual Computing Symposium (CVCS), 2015*, Gjovik, Norway, 25-26 August 2015, pp. 1-6. <https://www.its.bldrdoc.gov/publications/2682.aspx>

APPENDIX A INSTRUCTIONS

"Thank you for coming in to participate in our study. The purpose of this study is to gather individual perceptions of the quality of short video files. This will help us to understand video coding and transmission systems.

In this experiment you will be presented with a series of short clips. Each time a clip is played, you will be asked to judge the quality of the clip. A rating scale will appear on the screen and you should use the mouse to select the rating that best describes your opinion of the clip. Please rate the quality of video rather than the subject matter. For example, if the clip shows clowns and you dislike clowns, try to ignore that. Everything else you see should be considered. After you have clicked on one of the options, click on the "Vote" button to automatically record your response to the hard drive.

Do not worry about somehow giving the wrong answer; there is no right or wrong answer. Everyone's opinion will be slightly different. We simply want to record your opinion.

Two extra options are available: human error and computer error. Select "human error" if you cannot rate the clip. For example, you were distracted and did not pay attention to the video. This may happen. Select "computer error" if our playback system has a problem. For example, you see a flash of a different video at the beginning or end of the video. You should not encounter any computer errors. We once encountered a very rare error where the system repeatedly played two clips. If this happens, let me know.

Do you have any questions before we begin?

<…>

We will start with 12 clips while I am here with you. This is the starting screen for the test. Press the "start session" button on the menu bar to return to this starting screen. The software will not remember your user number, so you will need to enter that number for each session. Next to the user number is a pull-down list where you select the session name. After the training session, you will select sessions randomly using these cards.

Press F11 to enter full screen mode, then press the "submit" button to begin."

<after training>

"This experiment consists of 8 sessions, each focusing on a different topic. Our intention is that each session should feel like you selected a different video on the internet. The sessions are self-paced and will last about 11 to 17 min.

One session's theme is public safety. This includes simulated wounds and guns firing simulated teargas. If the content will bother you, please skip this session. You may also stop the public safety session at any time by pressing F11 to exit full screen mode, selecting "Start Session", and choosing your next session.

After each session finishes, the computer will tell you that the section is finished. Take a break. I will be here with you throughout the experiment."

APPENDIX B FOOTAGE ATTRIBUTION, LICENSE TERMS, AND EDITING ERRORS

Table B-1 lists the attribution for each video footage. Some the video footage cannot be distributed; these are marked "internal only." In all other cases, ITS has the right to distribute the footage on the Consumer Digital Video Library (CDVL, www.cdvl.org) for research and development purposes. See the CDVL website for the user agreement. A clarification in plain English is available at https://www.its.bldrdoc.gov/resources/video-quality-research/video-footage.aspx. Additional constraints apply to some of the footage, such as a document that must be kept with the footage or an obligation to identify the copyright holder in publications. These constraints are listed in column "Attribution and Copyright." Column "SRC Code" is the two digit footage attribution code used in the video file naming convention.

The videos in the **its4s** dataset were edited from the earliest version available to ITS, which we will refer to as the initial footage. Column "Initial Format" of Table B-1 identifies the resolution and frame rate of this initial footage. If the footage was filmed or commissioned to be filmed by ITS, then the initial footage was the recording format of a professional camera. If the footage was contributed by another organization or edited from outtakes, then the initial footage was edited and perhaps format converted before distribution to ITS.

Tables B-2 through B-10 identify the RSRC sets within each session. Column "RSRC" is a single letter code used in the video file naming convention. Columns "SRC Name" and "SRC Code" correspond to the attribution information in Table B-1. Column "RSRC Description" briefly describes the type of content in each RSRC pool. Each session uses different definitions for the RSRC codes A through Z, with the exception of "U" which is always used as a catch-all for unique scenes (i.e., where no related content exists). There are two bookkeeping errors, where one RSRC code is used for two different types of content in a single session. These are marked with footnotes.

All video files are named according to the naming convention "**A_B-CsrcD_E**" where:

- "A" is the session name (see column "Session" in Table 2)
- "B" is a three digit number that uniquely identifies that one video sequence within one session
- "C" is the session code (see column "Code" in Table 2).
- "D" is the RSRC code (see column "RSRC" in the specific table for that session).
- "E" is the HRC name. Notice that "SRC" is used instead of "original" within the file names.

For example, video file "Broadcast_001-Bsrc11A_1256K" is from broadcast session, RSRC "A", HRC 1256K, taken from source content 11 (Tears of Steel). The file number (001) is required to uniquely identify each file, because an RSRC set may associate two or more clips with the same HRC. The Everglades session provides an example. The RSRC "K" (animal shots)

contains four original videos: “Everglades_015-Esrc33K_SRC”, “Everglades_067-Esrc33K_SRC”, “Everglades_074-Esrc33K_SRC” and “Everglades_076-Esrc33K_SRC”.

Table B-11 identifies the editing errors, all of which were either two or three frames at the beginning of the clip that were from a different content. Note that subjects were asked to note any editing errors, but no subject detect the presence of these editing errors. See Section 4 for more information.

Table B-1. Footage Attribution.

SRC Name	Attribution and Copyright	SRC Code	Initial Format
Ancient Thought	Cable Labs (http://www.cablelabs.com/resources/4k/) This website is dedicated to providing next generation video content for free under the Creative Commons License.	14	4K 24fps
Animals	© Bennet-Watt HD productions (www.bennett-watt-hd-stock-footage-library.com)	20	1080i 29.97fps
Beekeepers	© Bennet-Watt HD productions (www.bennett-watt-hd-stock-footage-library.com)	29	1080i 29.97fps
Big Buck Bunny	Open movie “Big Buck Bunny” © copyright 2008, Blender Foundation, www.bigbuckbunny.org	12	1080p 24fps
Boxing	Boxing promotional video commissioned by NTIA/ITS and produced by Fireside Productions (www.firesideproduction.com). This footage was made possible by Touch ‘Em Up Boxing.	07	1080i 29.97fps
Cattle	© Bennet-Watt HD productions (www.bennett-watt-hd-stock-footage-library.com)	31	1080i 29.97fps
Cityscape	Footage commissioned by NTIA/ITS in 2013, and filmed by Fireside Productions (www.firesideproduction.com). These sequences were filmed on a Red One 4K camera.	30	4K 24fps
Eldorado	Cable Labs (http://www.cablelabs.com/resources/4k/) This website is dedicated to providing next generation video content for free under the Creative Commons License.	34	4K 23.976fps
El Fuente	Footage made available by NETFLIX. For more information, please refer to document titled “Netflix El Fuente Assembly Instructions” at http://www.cdvl.org/documents/index.php.	02	4K 23.976
Emergency Telemedicine	This emergency medical service (EMS) footage was choreographed, filmed, and contributed through the cooperative efforts of the National Association of State EMS Officials, the Western Eagle County Ambulance District of Colorado, Tristate CareFlight 15, the General Eagle Fire Protection District, Big Steve's Towing, NTIA/ITS and the Public Safety Communication Research (PSCR) laboratory (www.pscr.gov). The goal was to promote research and development into the use of video for emergency medical response.	23	1080i 29.97fps
Everglades	Video created by Colorado State University Journalism and Media Communication Professor Greg Luft	33	1080p 29.97fps
Fishin’ Florida	© Catamount Productions (www.catamountvideo.com) Taken from a fully edited sequence available on CDVL under the key words “NTIA Fishin Florida.”	19	1080p 29.97fps

SRC Name	Attribution and Copyright	SRC Code	Initial Format
Flamenco	Flamenco dance sequence that was commissioned by NTIA/ITS and filmed by Fireside Productions (www.firesideproduction.com). The edited sequence can be found on CDVL by doing a key word search for "flamenco." This sequence was made possible by Flamenco with Natalia.	01	1080i 29.97fps
The Foot	Four-minute music video of "High Design," an original song by the band, the Foot. It was released on their debut album Primary Colors. This music video was commissioned by NTIA/ITS and filmed by Fireside Productions (www.firesideproduction.com). The edited sequence can be found on CDVL by doing a key word search for "the foot."	03	1080i 29.97fps
Football Crowds	Crowds of people in a sports stadium, commissioned by NTIA/ITS and filmed by Fireside Productions (www.firesideproduction.com). This footage was filmed as part of the Public Safety Communication Research (PSCR) project, a joint endeavor of NTIA/ITS and NIST.	22	2080p 29.97fps
Geese	Video created by Colorado State University Journalism and Media Communication Professor Greg Luft	33	1080p 29.97fps
Great Wall	"Great Wall" from the Technicolor 3D video sequences, Copyright © 2013, Technicolor. These sequences can only be used for the purpose of research, and for the purpose of developing and testing technology standards. These sequences cannot be used for tradeshows or commercial purposes. See document "Technicolor_3D_videos_agreement.docx" for more details. That document (Technicolor_3D_videos_agreement.docx) must be distributed with the video sequences.	25	1080p 29.97fps
Internal Only	ITS only has rights to use this footage internally. These portions of the **its4s** dataset cannot be shared.	06 26	various
ITS Promotional	Footage describing the function of NTIA/ITS. This footage was created by students of the Colorado State University.	27	720p 60fps
Kenpo	Commissioned by NTIA/ITS and filmed by Interface Media Group.	04	Stereoscopic 3D pairs of 1080p 29.97fps
La Jolla	Commissioned by NTIA/ITS and filmed by Crystal Pyramid Productions (http://www.sandiegovideoproduction.com/).	32	1080i 29.97fps (1440 × 1080)
Lifting Off	Cable Labs (http://www.cablelabs.com/resources/4k/) This website is dedicated to providing next generation video content for free under the Creative Commons License.	08	4K 23.976fps
Liquid Assets	This footage was taken by Liquid Assets.tv productions during the filming of The travel show Into The Drink. For more info visit www.intothedrink.tv	16	1080i 29.97fps
Liquid Assets	(see above)	17	1080p 29.97fps

SRC Name	Attribution and Copyright	SRC Code	Initial Format
Mock Prison Riots	This footage was supported by Award No. 2009-IJ-CX-K016, awarded by the National Institute of Justice, Office of Justice Programs, U.S. Department of Justice. The opinions, findings, and conclusions or recommendations expressed in this footage are those of the author(s) and do not necessarily reflect those of the Department of Justice. The footage was also an initiative of the West Virginia High Technology Consortium Foundation. Information on the program can be found at: http://mockprisonriot.org. The filming was commissioned by NTIA/ITS and produced by Fireside Productions (www.firesideproduction.com), as part of the Public Safety Communication Research (PSCR), a joint endeavor of NTIA/ITS and NIST.	24	1080i 29.97fps
NASA	Public domain footage from NASA. See http://www.nasa.gov/multimedia/hd/HDGalleryCollection_archive_2.html. These training sequences include "Earth in HD: Stunning views of the home planet, from NASA's unique perspective in orbit"; "Arthur Christmas"; "NASA's SDO Captures Stunning 4K View of April 17 Solar Flare" from the NASA's Goddard Space Flight Center (http://svs.gsfc.nasa.gov/12224); "Moon Phase and Libration, 2015" (http://svs.gsfc.nasa.gov/4236); "Spacex Launch"; "HD Earth Views"; "RBSP Launch"; and "Grail Launch"	35	various
Saturated Feathers	Filmed by ITS on a Panasonic P2HD AJ-HPX3000G with a Fujinon TV Lens HA22x7.8 BERM-M48. This camera records in H.264 intra-frame coding at 100 Mbps. The camera was loaded with the "Musikvid" settings, which gives the sequence saturated colors. The vertical bars mimic the ITU Popple sequence.	27	1080p 29.97fps
Simulated News	Simulated news sequences that were commissioned by NTIA/ITS in 2004, and filmed by Fireside Productions (www.firesideproduction.com). These 24 sequences can be found on CDVL by doing a key word search for "simulated news" matching all words in the title only. Videographer was instructed to emphasize fast motion (e.g., a vehicles crossing the screen in one second, simultaneous zoom and pan).	13	1080i 29.97fps
Skateboarding	Cable Labs (http://www.cablelabs.com/resources/4k/) This website is dedicated to providing next generation video content for free under the Creative Commons License.	10	4K 23.976fps
Snowfall	Filmed by ITS on a Panasonic P2HD AJ-HPX3000G with a Fujinon TV Lens HA22x7.8 BERM-M48. This camera records in H.264 intra-frame coding at 100 Mbps. Footage depicts falling snow. CDVL has several videos edited from this footage, like "NTIA Colorado blue spruce (1e)".	28	1080p 25fps

SRC Name	Attribution and Copyright	SRC Code	Initial Format
SVT	© Sveriges Television AB (SVT). Individuals and organizations extracting sequences from the SVT archive agree that the sequences and all intellectual property rights therein remain the property of Sveriges Television AB (SVT), Sweden. These sequences may only be used for the purpose of developing, testing and presenting technology standards. SVT makes no warranties with respect to the materials and expressly disclaim any warranties regarding their fitness for any purpose.	15	1080p 50fps
Tears of Steel	Open movie "Tears of Steel" © Blender Foundation, mango.blender.org	11	1080p 24fps
TUM	COPYRIGHT NOTICE The TUM Multi Format Test Set is licensed under a Creative Commons Attribution-NonCommercial-ShareAlike 3.0 Germany License (http://creativecommons.org/licenses/by-nc-sa/3.0/de/deed.en). The use of the TUM Multi Format Test Set in any publication shall be attributed as: Technische Universitat Munchen, Institute for Data Processing. (2011) TUM Multi Format Test Set. [Online]. Available: http://www.ldv.ei.tum.de/videolab	05	720p 25fps, 1080p 25fps
Unspoken Friend	Cable Labs (http://www.cablelabs.com/resources/4k/) This website is dedicated to providing next generation video content for free under the Creative Commons License.	09	4K 23.976fps
Waves	© WideEye productions (www.wideeye.tv)	18	1080p 24fps

Table B-2. Broadcast Session Related Source Sequence (RSRC) Descriptions

RSRC	SRC Name	SRC Code	RSRC Description
A	Tears of Steel	11	Action movie
B	Tears of Steel	11	Detailed backdrop
C	Tears of Steel	11	Group of people talking
D	Tears of Steel	11	Pan and zoom
E	Tears of Steel	11	View through a firearm scope
F	Tears of Steel	11	Face seen close up
G	Tears of Steel	11	Movie credits
H	Big Buck Bunny	12	Animated movie
I	Big Buck Bunny	12	Scrolling movie credits
J	Simulated News	13	News, low motion
K	Simulated News	13	News, camera moving
L	Simulated News	13	News, fast movement
M	Ancient Thoughts	14	Candle in the dark
N	Ancient Thoughts	14	Monk with candles
O	SVT	15	Spotlight with confetti
U	Ancient Thoughts	14	Unique: disparate content

Table B-3. Chance Session Related Source Sequence (RSRC) Descriptions

RSRC	SRC Name	SRC Code	RSRC Description
A	Great Wall	25	Great wall of china
B	Internal Only	26	Beach adventure
C	Internal Only SVT	26 15	City pan
D	Internal Only	26	Mime in a mall
F	Saturated Feathers	27	Feathers and a spinning cage with saturated colors
G	ITS Promotional	27	Sketch episode: video animated from drawings
H	Internal Only	26	Helicopter
I	Snowfall	28	City views with heavy falling snow
J	SVT	15	Crowd of people running in a race
K	Beekeepers	29	Zoom on grass
L	Beekeepers	29	Tiny bees flying erratically
M	Beekeepers	29	Medium view of beehive
N	Beekeepers	29	Close view of beehive
O	Cityscape	30	Bus approaching at an angle
P	Cityscape	30	Children playing
Q	Cityscape	30	Flowers with a blurred backdrop
R	Cityscape	30	Kayaking with erratic water movements
U	TUM Beekeepers Cityscape	05 29 30	Unique: disparate content

Table B-4. Everglades Session Related Source Sequence (RSRC) Descriptions

RSRC	SRC Name	SRC Code	RSRC Description
F	Everglades	33	Nearly still nature shot
G	Everglades	33	View from a moving boat with a focused backdrop
H	Everglades	33	Distant scenery over water
I	Everglades	33	Close shot with moving water
J	Everglades	33	Nature shot with random motion
K	Everglades	33	Animal shots
L	Everglades	33	Wading birds
M	Everglades	33	People and props in a moving boat
U	Everglades	33	Unique: disparate content

Table B-5. Music & Mexico Session Related Source Sequence (RSRC) Descriptions

RSRC	SRC Name	SRC Code	RSRC Description
A	Flamenco	01	Wide view of the dances and musicians
B	Flamenco	01	Attention split between musicians and dancers' footwork
C	Flamenco	01	Close shot of musicians
C[5]	El Fuente	02	Dancers footwork
D	Flamenco	01	Pan
E	Flamenco	01	Object moves quickly across the screen
F	Flamenco	01	Close up of a face
G	Flamenco	01	Guitar fingering
H	El Fuente	02	Talker, subtle pan
I	El Fuente	02	Dim lighting with camera movement
J	El Fuente	02	Crowded street
K	El Fuente	02	Rain falling
L	El Fuente	02	River boat
M	El Fuente	02	Time lapse
N	El Fuente	02	Night shot
O	El Fuente	02	Noise from dim lighting
P	The Foot	03	Singer's face with hair bouncing
Q	The Foot	03	Camera movement
R	The Foot	03	Split attention
S	The Foot	03	Guitar fingering
T	The Foot	03	Crowd dancing
U	Flamenco El Fuente The Foot	01 02 03	Unique: disparate content
V	The Foot	03	Whole music video, with pan or crowd in foreground

Table B-6. Nature Session Related Source Sequence (RSRC) Descriptions

RSRC	SRC Name	SRC Code	RSRC Description
A	Cattle	31	Cattle herd, close up view
B	Cattle	31	Scenic cattle herd
C	Cattle	31	Cattle and dogs playing
E	Everglades	33	Buildings and equipment
F	Everglades	33	Beach, boats and waves
G	Everglades	33	Moving boat, seen from shore

[5] These two unrelated content types have the same RSRC code "C" due to an editing error.

RSRC	SRC Name	SRC Code	RSRC Description
H	Geese	33	Canadian geese swimming
I	Geese	33	Canadian geese flying
J	Geese	33	Canadian geese walking
K	Geese	33	Swimming close up
L	Geese	33	Sunset over water
M	Geese	33	Snowing
N	Geese	33	Geese taking off from snow
O	Geese	33	Geese on a white backdrop
P	Geese	33	Feeding geese
Q	Geese	33	Head in focus, blurred background
R	Eldorado	34	Mountain time lapse
S	Eldorado	34	Mountain wilderness pan
T	Eldorado	34	Rain
U	Animals Geese Eldorado	06 33 34	Unique: disparate content
V	Eldorado	34	Close up with blurred backdrop, still
W	Eldorado	34	Close up, camera moving forward
X	Eldorado	34	Walking with shoulder cam in mountains
D	La Jolla	32	Seabirds, seals and sea lions

Table B-7. Ocean Session Related Source Sequence (RSRC) Descriptions

RSRC	SRC Name	SRC Code	RSRC Description
A	Liquid Assets	16 17	Underwater
B	Liquid Assets	16 17	Ocean surface
D	Liquid Assets	17	Volcano at night
E	Liquid Assets	16 17	Underwater, many fish
F	Liquid Assets	17	Divers
G	Liquid Assets	17	Manta rays with numerous bubbles
H	Waves Animals	18 20	Ocean sunset
I	Waves Animals La Jolla	18 20 21	Peaceful waves
J	Waves La Jolla	18 21	Dramatic waves
K	Fishin' Florida	19	Fishing promotional
U	Waves	18	Unique: disparate content

Table B-8. Public Safety Session Related Source Sequence (RSRC) Descriptions

RSRC	SRC Name	SRC Code	RSRC Description
A	Football crowds	22	Football crowds, zoom 1 (wide view)
B	Football crowds	22	Football crowds, zoom 2
C	Football crowds	22	Football crowds, zoom 3
D	Football crowds	22	Football crowds, zoom 4
E	Football crowds	22	Football crowds, zoom 5
F	Football crowds	22	Football crowds, zoom 6 (close view)
G	Emergency telemedicine	23	Miscellaneous EMS footage
H	Emergency telemedicine	23	Wounds
I	Emergency telemedicine	23	Burn patient
K	Emergency telemedicine	23	Vehicle wreck extraction
L	Emergency telemedicine	23	Cardiac arrest
M	Mock Prison Riots	24	Riot with dim lighting and smoke
N	Mock Prison Riots	24	Wide view of a cell block
O	Mock Prison Riots	24	Close view of a riot
P	Mock Prison Riots	24	Simulated bodycam
Q	Mock Prison Riots	24	Riot with full sun

Table B-9. Sports Session Related Source Sequence (RSRC) Descriptions

RSRC	SRC Name	SRC Code	RSRC Description
A	Kenpo	04	Kenpo kata forms (i.e., choreographed martial arts patterns)
A[6]	TUM	05	Downhill skiing
B	TUM	05	Soccer
D	Animals Internal Only	06	Horse race
E	Boxing	07	Boxing, wide view of entire gym
F	Boxing	07	Close view of face
G	Boxing	07	Two people talking
U	Boxing Lifting Off Unspoken Friend Skateboarding	07 08 09 10	Unique: disparate content
H	Boxing	07	Close view, still
I	Boxing	07	Close view from shoulder camera
J	Boxing	07	Medium view from shoulder cam

[6] These two unrelated content types have the same RSRC code "A" due to an editing error.

RSRC	SRC Name	SRC Code	RSRC Description
K	Boxing	07	Still life
L	Lifting Off	08	Hot air balloon launch
M	Unspoken Friend	09	Fence lines
N	Unspoken Friend	09	Hooves spraying gravel, close view
O	Unspoken Friend	09	Horse running
P	Skateboarding	10	Skateboard store
Q	Skateboarding	10	Riding a skateboard
R	Skateboarding	10	Riding a skateboard, from shoulder camera
T	Skateboarding	10	Evening time lapse

Table B-10. Training Session Related Source Sequence (RSRC) Descriptions

RSRC	SRC Name	SRC Code	RSRC Description
A	NASA	35	NASA launch
B	NASA	35	NASA computer animation
C	NASA	35	NASA moon phases

Table B-11. Editing Errors

Video File Name	Extra Frames at Start
Broadcast_029-Bsrc11A_2340K	3
Broadcast_031-Bsrc11F_1732K	3
Broadcast_048-Bsrc12H_0512K	3
Broadcast_049-Bsrc12H_1256K	3
Chance_045-Csrc28I_1732K	3
Chance_048-Csrc15C_0951K	3
Chance_053-Csrc15J_1256K	3
Music&Mexico_052-Msrc02L_1256K	3
Ocean_010-Osrc16A_SRC	2
Ocean_087-Osrc19K_0951K	3
PublicSafety_083-Psrc24P_SRC	3
Sports_011-Ssrc05C_1732K	3
Sports_012-Ssrc05C_2340K	3
Sports_013-Ssrc05C_0951K	3
Sports_014-Ssrc05B_0512K	3
Sports_015-Ssrc05B_1256K	3
Training_002-Tsrc35A_SRC	2

Video File Name	Extra Frames at Start
Training_003-Tsrc35A_0512K	3
Training_006-Tsrc35A_0951K	3
Training_012-Tsrc35C_SRC	2

NTIA FORM 29
(4-80)

U.S. DEPARTMENT OF COMMERCE
NATIONAL TELECOMMUNICATIONS AND INFORMATION ADMINISTRATION

BIBLIOGRAPHIC DATA SHEET

1. PUBLICATION NO.	2. Government Accession No.	3. Recipient's Accession No.
4. TITLE AND SUBTITLE ITS4S: A Video Quality Dataset with Four-Second Unrepeated Scenes		5. Publication Date February 2018 6. Performing Organization Code NTIA/ITS.P
7. AUTHOR(S) Margaret H Pinson		9. Project/Task/Work Unit No. 3141012-300
8. PERFORMING ORGANIZATION NAME AND ADDRESS Institute for Telecommunication Sciences National Telecommunications & Information Administration U.S. Department of Commerce 325 Broadway Boulder, CO 80305		10. Contract/Grant Number.
11. Sponsoring Organization Name and Address National Telecommunications & Information Administration Herbert C. Hoover Building 14th & Constitution Ave., NW Washington, DC 20230		12. Type of Report and Period Covered

14. SUPPLEMENTARY NOTES

15. ABSTRACT (A 200-word or less factual summary of most significant information. If document includes a significant bibliography or literature survey, mention it here.)

This report describes the video quality subjective test **its4s**, including the experiment design and footage attribution. Subjective experiment **its4s** includes 813 unique video sequences, each four seconds in duration. No video sequences were repeated. The goals are (1) to provide insights into the optimal experiment designs for training no-reference (NR) metrics, and (2) to understand the impact of original video quality on mean opinion scores (MOS). Together these goals support the larger goal of progressing research on effective NR metrics. The dataset is freely available for research and development purposes.

16. Key Words (Alphabetical order, separated by semicolons)

subjective testing, image quality, video quality

17. AVAILABILITY STATEMENT	18. Security Class. (This report)	20. Number of pages
☒ UNLIMITED.	Unclassified	55
☐ FOR OFFICIAL DISTRIBUTION.	19. Security Class. (This page) Unclassified	21. Price:

NTIA FORMAL PUBLICATION SERIES

NTIA MONOGRAPH (MG)

A scholarly, professionally oriented publication dealing with state-of-the-art research or an authoritative treatment of a broad area. Expected to have long-lasting value.

NTIA SPECIAL PUBLICATION (SP)

Conference proceedings, bibliographies, selected speeches, course and instructional materials, directories, and major studies mandated by Congress.

NTIA REPORT (TR)

Important contributions to existing knowledge of less breadth than a monograph, such as results of completed projects and major activities.

JOINT NTIA/OTHER-AGENCY REPORT (JR)

This report receives both local NTIA and other agency review. Both agencies' logos and report series numbering appear on the cover.

NTIA SOFTWARE & DATA PRODUCTS (SD)

Software such as programs, test data, and sound/video files. This series can be used to transfer technology to U.S. industry.

NTIA HANDBOOK (HB)

Information pertaining to technical procedures, reference and data guides, and formal user's manuals that are expected to be pertinent for a long time.

NTIA TECHNICAL MEMORANDUM (TM)

Technical information typically of less breadth than an NTIA Report. The series includes data, preliminary project results, and information for a specific, limited audience.

For information about NTIA publications, contact the NTIA/ITS Technical Publications Office at 325 Broadway, Boulder, CO, 80305 Tel. (303) 497-3572 or e-mail ITSinfo@ntia.doc.gov.