HOW TO CHOOSE VIDEO SEQUENCES FOR VIDEO QUALITY ASSESSMENT

Margaret H. Pinson[1], Karen Sue Boyd[2], Jessica Hooker[1], and Kristina Muntean[1]

[1]National Telecommunications and Information Administration (NTIA)
[2]University of Colorado Boulder and University of Northern Colorado

ABSTRACT

This paper presents recommended techniques for choosing video sequences for subjective experiments. Subjective video quality assessment is a well understood field, yet scene selection is often limited by content availability. The Consumer Digital Video Library (www.cdvl.org) is a solution. Task oriented subjective testing is a newer field than entertainment oriented testing that may require a different approach to scene selection. We describe three different task-based investigations currently underway: performance requirements for public safety equipment, how quality affects comprehension of sign language over a video link, and how video affects oral comprehension over an audiovisual link. Recommendations for scene selection for two types of testing are given. The impact of experiment design will be considered. An example 1080i 29.97fps video sequence set is presented.

1. INTRODUCTION

Scene selection is an important component of video quality research. Selection should be based on video characteristics and the purpose of the experiment, not on personal preference or convenience. Task oriented subjective testing is a newer field, and raises new issues for scene selection. This paper first describes three task oriented subjective video quality assessment experiments. Then we describe guidelines for scene selection for both traditional, entertainment oriented tests and task oriented tests developed from over two decades of experience designing video quality subjective tests. Finally, we review experimental design considerations common to both types of subjective testing.

2. TASK ORIENTED SUBJECTIVE TESTING

The three task-based experiments under investigation by NTIA described here provide examples for the scene selection recommendations in section 4. Each task demonstrates a question about video quality that cannot be answered using a traditional subjective video quality test.

2.1. Video quality for public safety applications

Video is quickly emerging as an essential component of effective public safety communications. Example uses include evidence in criminal cases, aerial images of wildfires, highway traffic monitoring, and assessing accidents. As video technology has evolved, the equipment options have become increasingly complex. As a result, many public safety agencies lack the tools, support, and information they need to make informed video system purchasing decisions. Unbiased guidance is essential for practitioners to clearly articulate their video quality needs.[1] The task is to identify video system characteristics that allow public safety practitioners to perform their job.

2.2. Sign language comprehension

Remote communication using sign language requires a video link. The impact of video encoding on visual-gesture comprehension has been investigated in papers such as [2]-[4], the work of Gunnar Hellström,[1] and the Wireless Information Services for Deaf People on the Move project.[2]

Sign languages like American Sign Language (ASL) are visual-gesture languages entirely separate from their spoken counterparts. Thus, translation issues occur when going from sign language to spoken or written language, and vice versa. Sign languages consist of many different components and grammatical structures. Our test design focuses on five components:

- ***Vocabulary.*** Words require the use of one dominant hand, and two-handed symmetrical sign or two-handed non-symmetrical sign. Signs are structured and organized by four dominant articulator parameters: handshapes, palm orientation, location, and movement.
- ***Finger Spelling.*** Words can be spelled out using handshapes. In ASL, for instance each letter is

[1] Gunnar Hellström, "Quality measurements on video communication for sign language," (unpublished), Omnitor, Stockholm, Sweden, 1997.
[2] "WLAN Trials for sign language conversation," (unpublished), WISDOM, Jun. 2003.

associated with specific positioning of one hand and words can be spelled as quickly as five to eight letters per second.

- ***Non-manual Markers (NMM).*** Both grammatical structure and additional information are expressed through movements of eyes, eyebrows, mouth, tongue, head, neck, etc. For example, a signer might simultaneously mouth a particular word in English (i.e., without sounding out the word), while signing a word which is only available in sign language.
- ***Gestures.*** Hand or arm movements not considered part of sign language vocabulary may be used to supplement communication. For example, a gesture can show the signer's reaction to a large number of attendees, or indicate the signer is trying to remember what to say.
- ***Spatial and Role Shifting.*** When describing multiple people, places or things, each object is given its own position by pointing to a particular location in space. To distinguish between two or more people conversing, the body shifts, particularly the shoulder, head, and eyes.

The task is to identify video system characteristics that allow people full comprehension of sign language, without requiring the signer to change his or her behavior.

2.3. Speech perception through lip-reading

Some people with hearing loss comprehend speech through lip-reading. Accurate comprehension through lip-reading alone is difficult, because many of the speech sounds look the same visually (*e.g., /p/, /b/, and /m/*). At best, only about 30% of speech is clearly visible on the lips. There is an interaction between hearing and vision in speech perception among people with normal hearing. This can be demonstrated through the perceptual phenomenon known as the McGurk effect [5]. The task is to quantify the added benefit of video in the ability of people with hearing loss to perceive speech (i.e., comparing audio only stimuli with audiovisual stimuli).

3. ENTERTAINMENT ORIENTED SCENE SELECTION

Entertainment oriented subjective video quality tests try to represent a wide range of entertainment content in a scene pool containing perhaps 8 to 10 clips. Naturally this is impossible, but approaching this ideal improves accuracy of the test's results.

3.1. Consider content editing and camerawork

The impact of scene content editing and camerawork cannot be underestimated. Viewer instructions for subjective testing should include a statement such as: "Please do not base your opinion on the content of the scene or the quality of the acting." Yet ratings inevitably include both the clip's artistic quality and its technical quality. This is why subjective tests normally include the original video.

To illustrate this issue, consider the Video Quality Experts Group (VQEG) High Definition Television (HDTV) Test [6]. This international test produced six subjectively rated databases. The six scene pools were carefully selected and balanced by Margaret H. Pinson. During the selection process, all original scenes were judged to have a quality of "good" or better by a panel of video quality subjective testing experts.

Fig. 1 and Fig. 2 show sample frames from each of these six datasets. Fig. 1 shows samples from the original sequence with the highest mean opinion score (MOS). Fig. 2 shows samples from the original sequence with the lowest MOS. The average MOS drops from 4.7 in Fig. 1 to 4.1 Fig. 2. Consider the impact this might have on your data analysis!

The impact of editing and camerawork can be seen in these sequences. The Fig. 1 scenes contain more scene cuts, animation, vibrant colors and good scene composition. This adds visual interest and improves the esthetic appeal. By contrast, half of the Fig. 2 scenes contain no scene cuts. The Fig. 2 sequences contain a variety of minor problems that had a large cumulative impact on MOS, such as motion blur, analog noise, camera wobble, poor scene composition or boring content.

The videos from Fig. 1 and 2 are available on the Consumer Digital Video Library (CDVL, www.cdvl.org). While vqegHD6 cannot be redistributed, the vqegHD6 original video sequences shown in Fig. 1 and 2 can be seen by searching for titles "NTIA Green Bird" and "Common SRC 14".

3.2. Choose scenes that evenly span a wide range of coding difficulty

Easy-to-code scenes are widely available, because they are easy to shoot. Finding hard-to-code content is more challenging. To simplify the task of judging scene complexity, use an objective complexity metric such as SI and TI [7] or criticality [8]. Alternatively, the researcher can encode all video content at a low bit-rate, and classify each scene manually.

At a minimum, we recommend:

- Two clips that are very difficult to code (e.g., criticality [8] ≥ 3.5)
- Two clips that are very easy to code (e.g., criticality [8] ≤ 2.5)
- One high spatial detail clip (e.g., many small objects, SI [7] ≥ 200)
- One high motion clip (e.g., an object that moves across the screen in 1 second, TI [7] ≥ 60)

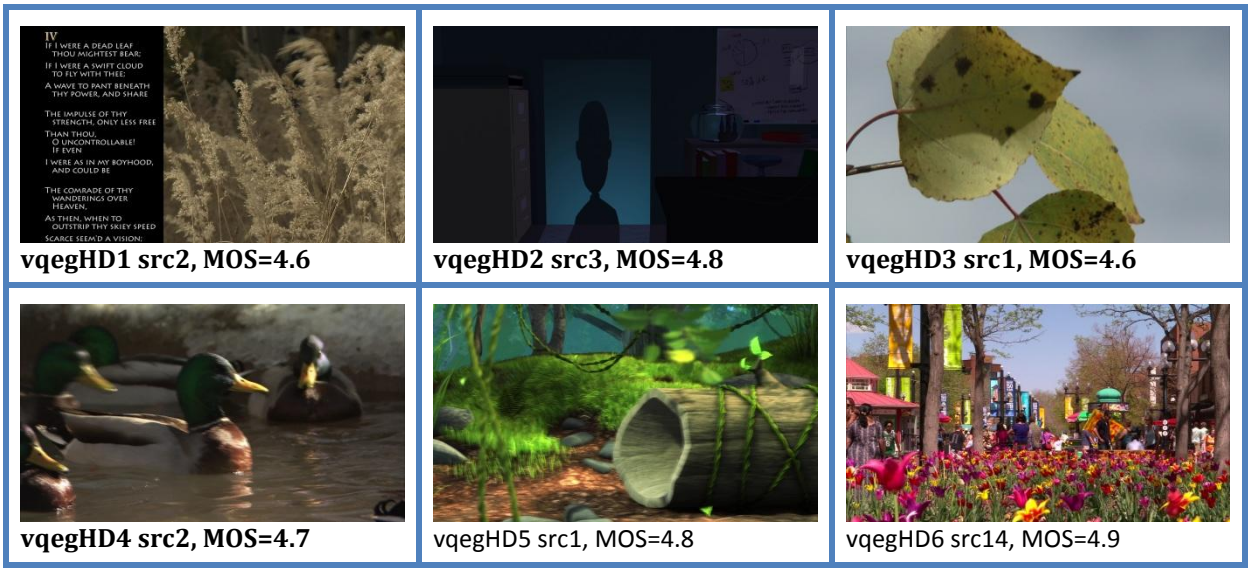

Figure 1. For each VQEG HDTV Phase 1 dataset, sample frame from original video sequence with highest MOS.

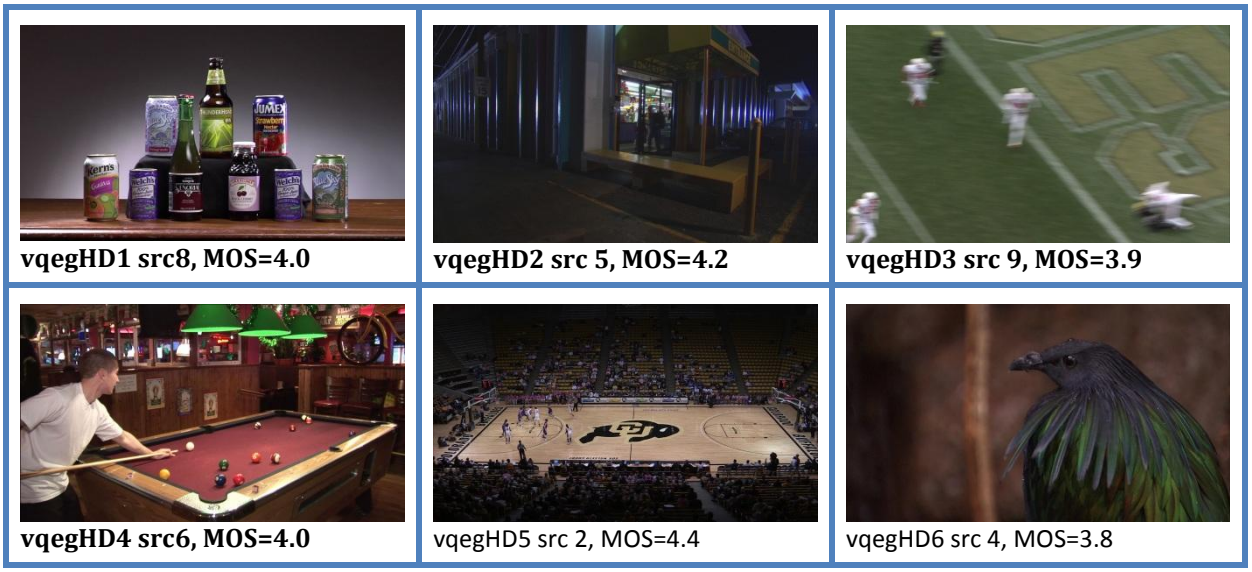

Figure 2. For each VQEG HDTV Phase 1 dataset, sample frame from original video sequence with lowest MOS.

Certain algorithmic flaws appear only in hard-to-code scenes, while others only appear in easy-to-code scenes. If the scene pool considers only easy-to-code scenes (or hard-to-code scenes), then system characteristic estimates will be flawed. Two examples are the relationship between bit-rate and quality, and the visual impact of transmission errors.

3.3. Consider frequency and placement of scene cuts

Scene cuts mask impairments from a few frames before the scene cut until about 0.25 seconds after it. Though coding algorithms can introduce a new group of pictures (GOP) in response to a scene cut, this affects the encoding. Scene cuts occur very frequently in movies and broadcast television; they do not typically occur in other applications such as videoconferencing or surveillance.

Scene cuts complicate subjective testing. The concern is that the encoded quality may be dramatically different, before and after the scene cut. The task of judging quality (or usability) thus becomes more difficult. Some researchers only select content that does not have scene cuts. This was the prevalent opinion expressed in VQEG and ATIS throughout the 1990s.

The problem is that these results may not fully represent user experiences. This is the prevalent opinion expressed in VQEG today. The lead author's preference regarding scene cuts is to select:

- About half of the clips with scene cuts
- One clip with rapid scene cuts
- About half of the clips without scene cuts

Note that the "differing quality" phenomenon is not unique to scenes with scene cuts. This also occurs spatially or temporally in continuously filmed content. Different parts may be better focused or intentionally blurred, relatively still or containing significant motion. Any of these variations will trigger quality differences that might make the subject's task more difficult.

3.4. Select scenes with unusual properties.

We learn the most from scenes that have unique traits. For example, consider a scene showing a close-up view of a person. Our test subjects know how people should look, move, and sound. This internal reference helps them notice the unnatural motion of a reduced frame rate. That low frame rate may be less obvious when watching a video of a machine.

The following scene traits can interact in unique ways with a codec or a person's perception. Our ideal scene pool includes all of these traits.

- Animation, graphic overlays, and scrolling text
- Repetitious or indistinguishable fine detail (e.g., gravel, grass, hair, rug, pinstripes)
- Sharp black/white edges
- Blurred background, with an in-focus foreground
- Night scene or dimly lit scene
- Ramped color (e.g., sunset)
- Water, fire or smoke (for unusual shapes and shifting patterns)
- Jiggling or bouncing picture (e.g., handheld camera)
- Flashing lights or other extremely fast events
- Action in a small portion of the total picture
- Colorful scene
- Small amounts of analog noise
- Multiple objects moving in a random, unpredictable manner

4. TASK ORIENTED SCENE SELECTION

Task oriented subjective video quality tests try to represent all types of video that might be used for a particular task. The guidelines in the previous section are less pertinent—except for coding difficulty, which is always important. The problem becomes how to span the full scope of a task, without introducing bias.

4.1. Consider the impact of context

Entertainment oriented subjective video quality tests ask people to rate the visual quality of video sequences. The experiment consists of a small set of sequences. Each sequence is impaired in a variety of ways, so that the ratings can be compared.

Task oriented subjective testing instead asks whether or not people can use a system for a specific purpose. Subjects are quizzed on their ability to comprehend what is occurring in a sequence—for example, identify the item held in a person's hand, or read a car's license plate number. The goal is to gain understanding of the relationship between video system characteristics and the ability of a person to use the video to perform one particular task.

If a sequence is re-used, subjects can answer comprehension-based questions from memory. One solution is to use each source sequence only once. This makes the data analysis difficult.

ITU-T P.912 offers a second solution: scenario groups. ITU-T P.912 combines the concept of the Modified Rhyme Test (MRT) for speech quality with a "Spot the Difference" image game. Each sequence group contains sequences that look alike except for small differences. For example, in [9], the same person walked by carrying a variety of objects—but the zoom, backdrop, etc., remained the same.

Unfortunately, obtaining entirely unbiased results from either approach is difficult, because people are very clever at picking up contextual clues. Using [9] again as an example, subjects might have notice unintended patterns in the objects (e.g., a uniquely colored object) or unintentional differences within the filming (e.g., cloud formations).

The same phenomenon occurs in speech perception sequences. Our experiment uses natural speech. Footage from one speaker naturally contains speech about related topics. Care must be taken when selecting speech segments to avoid providing contextual clues that will inadvertently enhance subjects' ability to comprehend other segments of that same speaker's footage.

4.2. Represent all behaviors.

Our visual-gesture experiments seek to understand how video quality affects comprehension of visual-gesture language. A limitation of some previous studies is that they focused on a sub-set of visual-gesture language. For example, [2] used vocabulary signs, and [3] used finger spelling. These experiments each focused on one of the five visual-gesture language elements listed in section 2.2, so the results may not generalize to other language elements.

Similarly, [4] used footage from one fluent signer. The problem is that each signer has a different signing system. The signer may have a fast or slow signing pace. The signer may make wide gestures or may keep them close to the torso. The signing system also influences how

often someone uses the different elements—for example, finger spelling may be rare or frequent, role shifting may or may not be used, and different types of NMM may be preferred.

To be accurate, audio experiments require a variety of different talkers. Similarly, task oriented subjective tests require a wide sampling of the task to be accomplished. Without this variety, the experiment's conclusions cannot generalize to the wider population of people performing that task.

4.3. Choose natural or artificial behavior.

The advantage of artificial stimuli is that the experimenter has full control over the stimuli. This eliminates uncontrolled variables, and strengthens the types of conclusions that can be reached. This is particularly important when examining immature technologies, such as stereographic 3D television and ultra-high definition television.

Audio subjective tests typically use sentences carefully crafted for that experiment. Speech quality tests prefer phonetically-balanced sentences, such as the Harvard balanced sentences. Rhyme tests use single words. Lip-reading experiments typically use single words or short sentences that can be easily remembered. The scenario groups for public safety experiments [9] and [10] each showed a person holding a variety of objects. Another example of artificial stimuli is an entertainment oriented video quality test that use only scenes without scene cuts.

The disadvantage is that artificial stimuli are inherently different from natural stimuli. People sound different when they are reading from a script. Spontaneous speech is filled with "um" and "ah," improper grammar, slurred words, sentences that never seem to end, and redundant information. In real conversations, we expect our listeners to use contextual clues to enhance their speech perception. For the speech perception through lip-reading task, our goal is to measure comprehension of the message—not accuracy recognizing individual sounds. Thus, spontaneous speech stimuli are desirable.

Artificial scenes cannot fully characterize every attribute of the stimuli you want to examine. Thus, you should start with artificial stimuli to control the scope of your problem and obtain initial results. At some point, it would be wise to confirm those findings using natural content—just to make sure you aren't missing something.

5. EXPERIMENT DESIGN

The goal of experiment design is to apply the scientific method to objectively answer a question. The critical issue is that scene choices must not bias the results.

5.1. Don't skimp on your scene total

Experiment design is always a compromise between the number of impairments, the number of scenes, and each subject's participation time. It is tempting to reduce the number of scenes (or subjects) so that the number of impairments can be increased. Resist this urge! It is impossible to demonstrate a wide range of behaviors if your scene pool has only two sequences.

5.2. Avoid over training by maximizing diversity

A common problem is selecting scenes from the same small pool of video content. This biases research results toward characteristics of those video sequences. Over training is a likely byproduct of small scene pools or reusing the same sequences in multiple experiments. Instead, we encourage you to find new sequences for each experiment.

CDVL provides free downloads of video clips for research and development purposes. CDVL is a repository of broadcast quality video content and subjectively rated databases. The goal is to foster research and development into consumer video processing and quality measurement.

5.3. Do scene selection on the device to be tested

Video quality subjective testing has traditionally involved uncompressed video played to broadcast quality monitors. This removed the effect of the video playback and monitor from the data, and helped us focus on video encoding, network transmission and video decoding.

That approach is impossible for mobile devices. These experiments must use compressed playback and lower quality monitors—and account for their impact on the subjective data. The computer used to view, select, edit and impair the video is probably a more powerful computer—perhaps a high end PC with a large monitor. Switching to the device under test will impact the appearance of your sequence. [11] We recommend that you always perform final scene selection on the device under test.

6. IMPACT

The impact of number of scenes on an experiment can be seen in Pinson *et al.* [12]. This article analyzes thirteen subjective experiments. Each explored the relationship between audio quality (***a***) and video quality (***v***), measured separately, and the overall quality of an audiovisual experience (***av***). One way to measure this is the Pearson Correlation between ***av*** and the cross term (***a*** x ***v***). Fig. 3 shows a histogram of these correlations, split by the number of scenes in the experiment: limited (one or two) and normal (five to ten). The former spans the range

[0.72..0.99], indicating that chance played a large role. The latter are tightly clustered, indicating a high degree of repeatability.

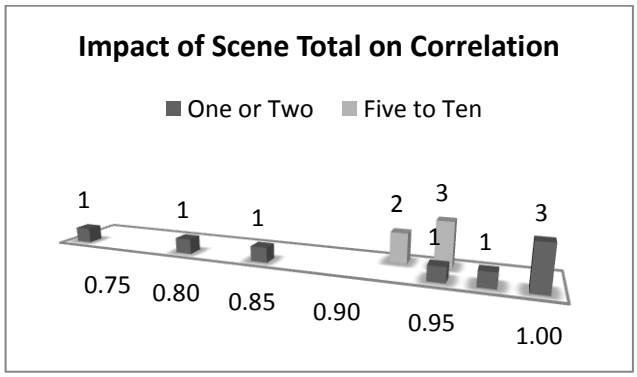

Figure 3. Histogram of the Pearson correlation from 13 subjective experiments. Each was designed to answer the same question. Data is divided by the number of scenes in the experiment.

Poor scene selection is difficult to detect during data analysis; good scene selection is more obvious. For example, Barkowsky et al. [13] analyzes a subjective test that investigated the QP parameter on the quality of H.264 encodings. This experiment's nine scenes were chosen using the criteria from Sections 3 and 5. Six scenes show similar QP/quality response curves, while the other three show unique behaviors. Without those three scenes, the accuracy of QP to predict quality would have been inflated.

7. CONCLUSION

The scene selection criteria in this paper were developed over two decades of subjective video quality experiments on a variety of topics. The topics examined included objective video quality model training and testing, comparisons of various codecs, analysis of different network error response strategies, and coding parameter optimization.

We will close by suggesting an example scene pool for entertainment experiments. This pool includes most of the characteristics recommend in this paper. These clips can be found on CDVL. They are 1920×1080, interlaced, 59.94 fields-per-second, and not edited for consistent length.

- "NTIA Flamenco Dancers Segment 1 frames 1 to 1000 of 12295"
- "NTIA Touch Em' Up Boxing Segment 4 frames 3001 to 4000 or 6047"
- "Bennet-Watt HD, Tramore Horse Racing from spectators angle"
- "Liquid Assets, greenschool"
- "NTIA Burn Close-up"
- "NTIA simulated news budget"
- "NTIA snowy day in the city (3a)"
- "NTIA The Foot music video Segment 4"

7. REFERENCES

[1] Public Safety Communications Research, www.pscr.gov/

[2] R. A. Foulds, "Biomechanical and perceptual constraints on the bandwidth requirements of sign language," *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, vol. 12, no. 1, 2004, p. 65-72.

[3] P. Heribanova *et. al.*, "Logatom intelligibility of single-handed finger alphabet", *ELMAR*, 2011, p. 71-74.

[4] F. M. Ciaramello *et. al.*, "Quality versus intelligibility: studying human preferences for American sign language video", *Image Processing Workshop (WNYIPW)*, Nov. 2010.

[5] http://video.its.bldrdoc.gov/videos/mcgurk/

[6] "Video Quality Experts Group: report on the validation of video quality models for high definition video content," Jun. 2010. Available: www.vqeg.org

[7] ITU-T Recommendation P.910, "Two criteria for video test scene selection," Geneva, 1994, section 6.3. Available: www.itu.int

[8] C. Fenimore *et al*, "Perceptual effects of noise in digital video compression," 140th SMPTE Technical Conference, Pasadena, CA, Oct. 28-31, 1998.

[9] "Recorded-video quality tests for object recognition tasks," DHS-TR-PSC-11-01, US Department of Homeland Security, Sep. 2011. Available: http://www.pscr.gov/outreach/safecom/vqips_reports/RecVidObjRecogn.pdf

[10] "Video quality tests for object recognition applications," DHS-TR-PSC-10-09, US Department of Homeland Security, Sep. 2010. Available: http://www.safecomprogram.gov/library/Lists/Library/Attachments/231/Video_Quality_Tests_for_Object_Recognition_Applications.pdf

[11] A. Catellier *et. al.*, "Impact of mobile devices and usage location on perceived multimedia quality," Quality of Multimedia Experience (QoMEX), Jul. 2012.

[12] M. Pinson *et. al.*, "Audiovisual quality components," *IEEE Signal Processing Magazine*, vol. 28, no. 6, Nov. 2011, p. 60-67.

[13] M. Barkowsky *et al.*, "Analysis of freely available subjective dataset for HDTV including coding and transmission distortions," *VPQM 2010*, Jan. 2010.