Film265: Report on Achieved Quality Improvements: Public Version

Document Information

Project Acronym	Film265
Contract Number	645500
Project Website	www.film265.eu
Contractual Deadline	30/06/2016
Nature	Report
Dissemination Level	Public
Authors	Sergio Sanz-Rodríguez (TU Berlin), Mauricio Álvarez-Mesa (TU Berlin)
Contributors	
Reviewers	Tilman Sheel (reelport)

Notices:
The research leading to these results has received funding from the European Community's Horizon 2020 Programme under grant agreement No. 645500.

Revision History

Comment	Version Number	Date
Final Version	1.0	30/06/2016
Minor Adjustments for Open Access	1.1	26/08/2016

Table of Contents

List of Acronyms

BD	Bjøntegaard Difference
BTC	Basic Test Cell
DCT	Discrete Cosine Transform
DSCQS	Double Stimulus Continuous Quality Scale
DSIS	Double Stimulus Impairment Scale
FIF	Frames-in-Flight
GOP	Group of Pictures
HD	High Definition
ITU	International Telecommunication Union
MOS	Mean Opinion Score
MSE	Mean Squared Error
PSNR	Peak Signal-to-Noise Ratio
QoS	Quality-of-Service
SD	Standard Definition
SIMD	Single Instruction Multiple Data
SSIM	Structural Similarity
VBR	Variable Bit Rate
VoD	Video-on-Demand
WPP	Wavefront Parallel Processing

1 Executive Summary

This document presents the results of the informal video quality subjective tests performed in the context of the Film265 project. In this project a new implementation of the HEVC/H.265 video coding standard has been developed that obtains higher quality and higher compression efficiency than previous codecs such as the H.264/AVC, and other implementations of the HEVC/H.265 encoder (such as the open source x265 encoder). An informal subjective analysis has been carried out in order to validate the results obtained in objective comparisons. In the context of the Film265 project it was important to have a test scenario that is very similar to the final use case of the Video-on-Demand (VoD) applications in the film industry. As a result, a test environment was created that allowed users to test the quality of several videos encoded with different encoders using a web playback platform. Although the test environment differs from those of rigorous formal video quality subjective tests, the results obtained in this test are closer to the final application.

For creating the content used in the evaluation, an enhanced video transcoder based on FFmpeg (a popular open source media transcoder) was used. This version of FFmpeg includes support for both H.264 (current encoder used in VoD services) and the new encoder developed in the project called TUB-H.265.

The main objective of the informal subjective test was of asses the visual experience obtained when using the new H.265 encoder compared to the H.264 encoder that is currently used in VoD applications for the film industry.

Informal subjective tests were performed with both regular VoD users and professionals in the video coding and production field. Opinions on the subjective experience and perceived differences between videos encoded with the two codecs where collected. Although these subjective tests were essentially informal, some recommendations reported in standardized methods for subjective quality assessment of videos have been taken into account and adapted for the purposes of the Film265 project. The main purpose has been to assess the quality gains of H.265 compared to H.264 in a web-based VoD environment.

The outcome of the subjective tests has been used as a feedback for improvements of the TUB-H.265 encoder, and also as a guideline for selecting appropriate quality-rate points for H.265 compressed videos that can cover different Qualities of Service (QoS).

2 User Feedback and Final Adjustment Encoding System

2.1 Description of Task

According to the Film265 Grant Agreement the description and objectives of this task are:

“Based on the first working prototype of the PicturePipe system with the new H.265 encoder and the new web H.265 player a set of tests and adjustments will be conducted. The tests will be done on the real production systems of Reelport, Cinando and LevelK comparing the existing and new encoding system. Feedback will be collected from the VoD providers about the experience with the new system, especially regarding the quality and bitrate. Based on the received feedback, adjustments will be made to the encoding system. This task includes an interactive process that will be performed until the end of the project with the objective of finding the best settings of encoding configuration parameters for the scope of applications of the VoD users."

2.2 Relevance of the Task to the Project

This task basically aims at confirming, from the subjective quality perspective, the codec comparison results obtained in previous development tasks. Although the TUB-H.265 encoder is able to achieve approximately 50% bitrate reduction for same Peak Signal-to-Noise Ratio (PSNR) compared to x264 when encoding HD (720p) and Full HD (1080p) movies, it is necessary to evaluate the final user's visual experience. According to some research publications (Ohm, 2012), the H.265 reference software achieves 40% bitrate reduction for the same PSNR, and up to 50% bitrate reduction for same subjective quality compared to H.264 when using test content. This information only gives a rough idea on how much subjective improvement the TUB-H.265 implementation can reach, but such an enhancement must be verified by means of a (at least informal) subjective quality test using real-life content.

Subjective tests have been conducted by expert and non-expert viewers in the video coding and production field, and the outcome of these tests has been used as a feedback for improvements of the encoder, also has been used for selecting the quality-rate points of the TUB-H.265 encoder integrated into PicturePipe for encoding the target VoD content.

2.3 Work Performed

2.3.1 Perceptual Quality Degradation in the Video Processing Chain

Since the video source is acquired by a video camera until it is displayed at the end-user side, a digital video signal goes through different stages in the video processing chain (see Figure 1) that can degrade the visual quality and, hence, the final user's visual experience. All the potential artefacts produced by the elements of the chain can be categorized into three classes: acquisition artefacts, compression artefacts, and transmission and playback artefacts.

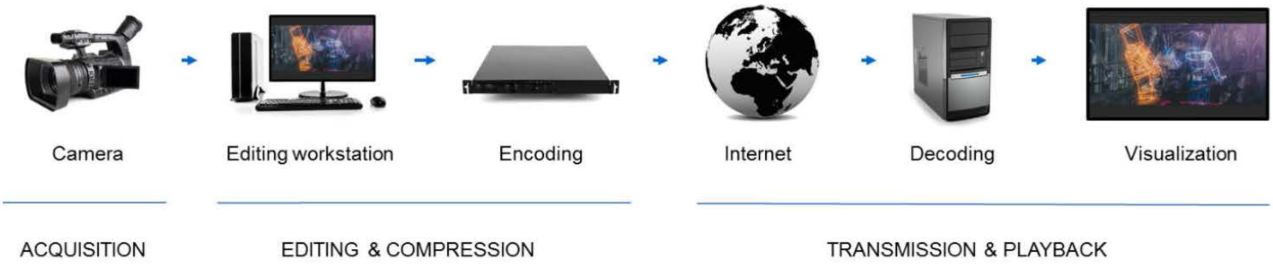


Figure 1. Video processing chain for transmission over the Internet

2.3.1.1 Acquisition Artefacts

A video camera consists of an optical system, a sensor that converts light into electrical signal, and additional processing circuity. Imperfections in these elements can create artefacts in the captured video. Optical imperfections result in effects such as defocusing in some parts of the picture, barrel distortion (the centre of the picture appears magnified), or vignetting distortion (light intensity is reduced from the centre of the picture to the periphery). Capture distortion includes interlaced scanning, spatial and temporal aliasing (A. Punchihewa, 2002).

Some amateur and semi-professional cameras may produce electrical noise, making the captured video grainy. Similar effect is produced when adding electronic gain in cameras to light dark scenes. Any type of camera noise has a negative impact on the final bitrate when compressing the video.

In the context of Film265 there is no control of acquisition artefacts and no way to mitigate them, since PicturePipe accepts any kind of video file, which is transcoded via FFmpeg.

2.3.1.2 Compression Artefacts

Modern hybrid video encoders such as HEVC/H.265 introduce distortion due to the quantization of the input video. These artefacts can be subjectively visible and can become very disturbing specially at low bitrates. Several types of compression artefacts can be distinguished (Unterweger, 2013), specifically: blocking, blurring, mosquito noise, ringing, and stair case. All of them are briefly described below:

Blocking

Blocking artefact is produced because of the division of the picture into blocks. In the case of H.265 the picture is partitioned in 64x64 blocks and then sub-partitioned into smaller square and/or rectangular blocks. These blocks are encoded separately and, as a result, the block boundaries in the reconstructed picture become visible in some cases. Although the most modern video coding standards include tools to mitigate such an effect, blocking artefact can be still visible at very low bitrates particularly in uniform areas (like the sky) and very detailed areas in motion (like the rough water in the sea).

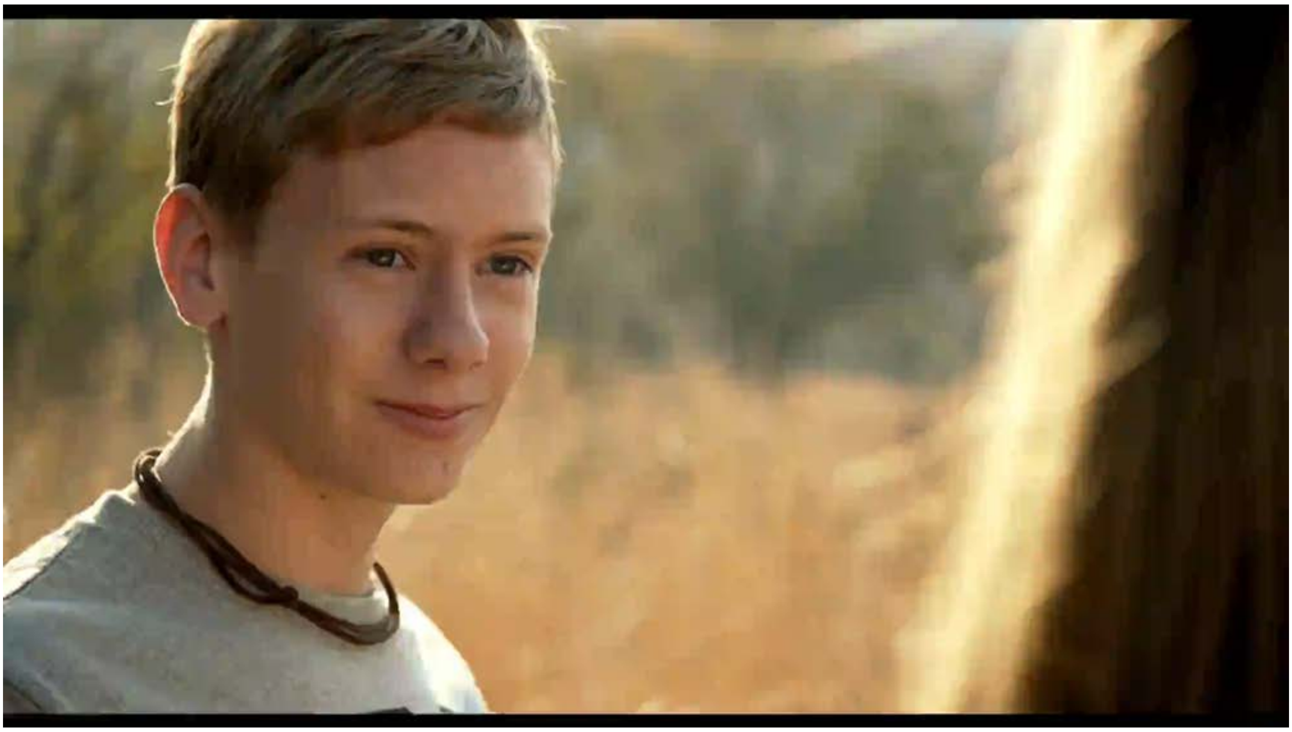


Figure 2. Example of blocking artefact

Banding

This effect is the result of the quantization on areas with a continuous gradation of colour tone or luminosity. As a consequence, a series of discrete steps or bands of colour appear in the picture, especially in uniform areas. Banding is specifically visible in 8-bit depth video compression.

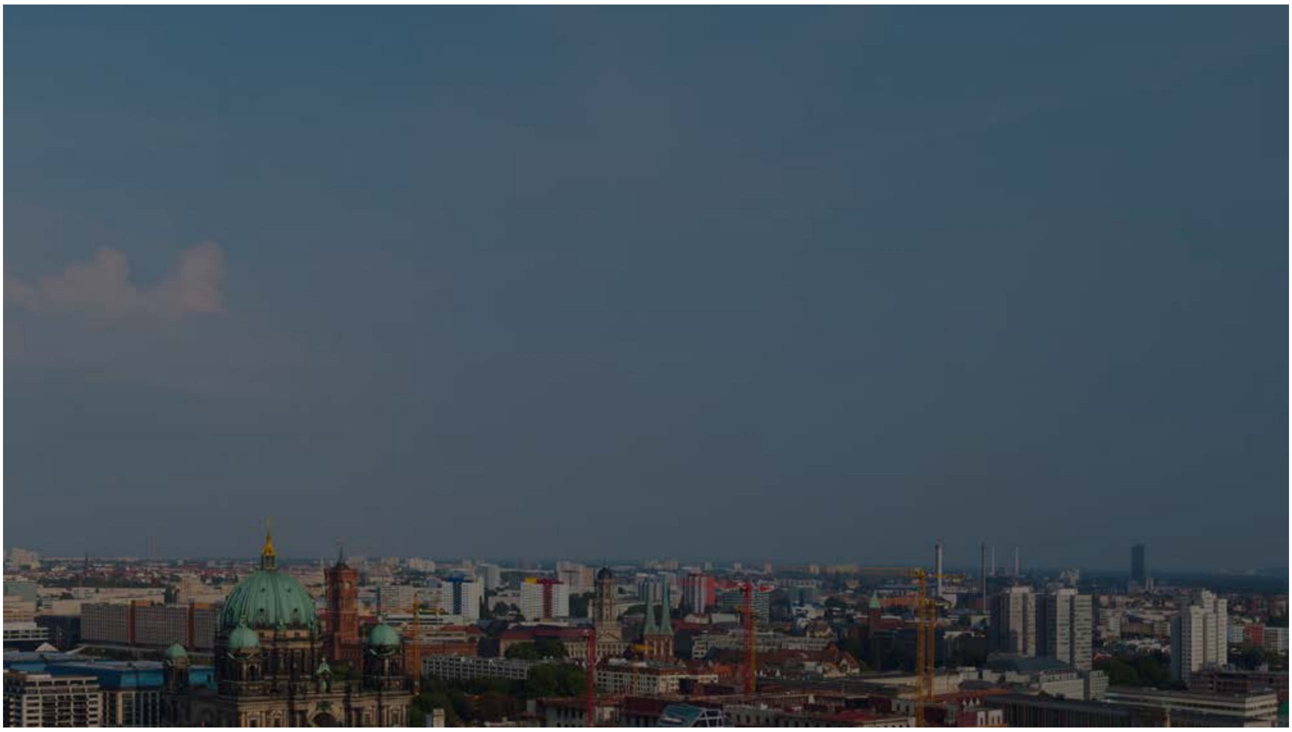


Figure 3. Example of banding artefact

Blurring

Generally speaking, blurring is produced when removing detail information in the picture by means of low-pass filtering. Blurring is also generated when blocks in the picture are strongly quantified. Such an effect is even enhanced by the operation of the deblocking filter in H.264 and H.265, which applies a smoothing operation just at the block edges to mitigate the blocking artefact. In the example below the detail in some parts of the picture has been completely removed by the encoder. As a consequence, the compressed signal is a blurred version of the original one.

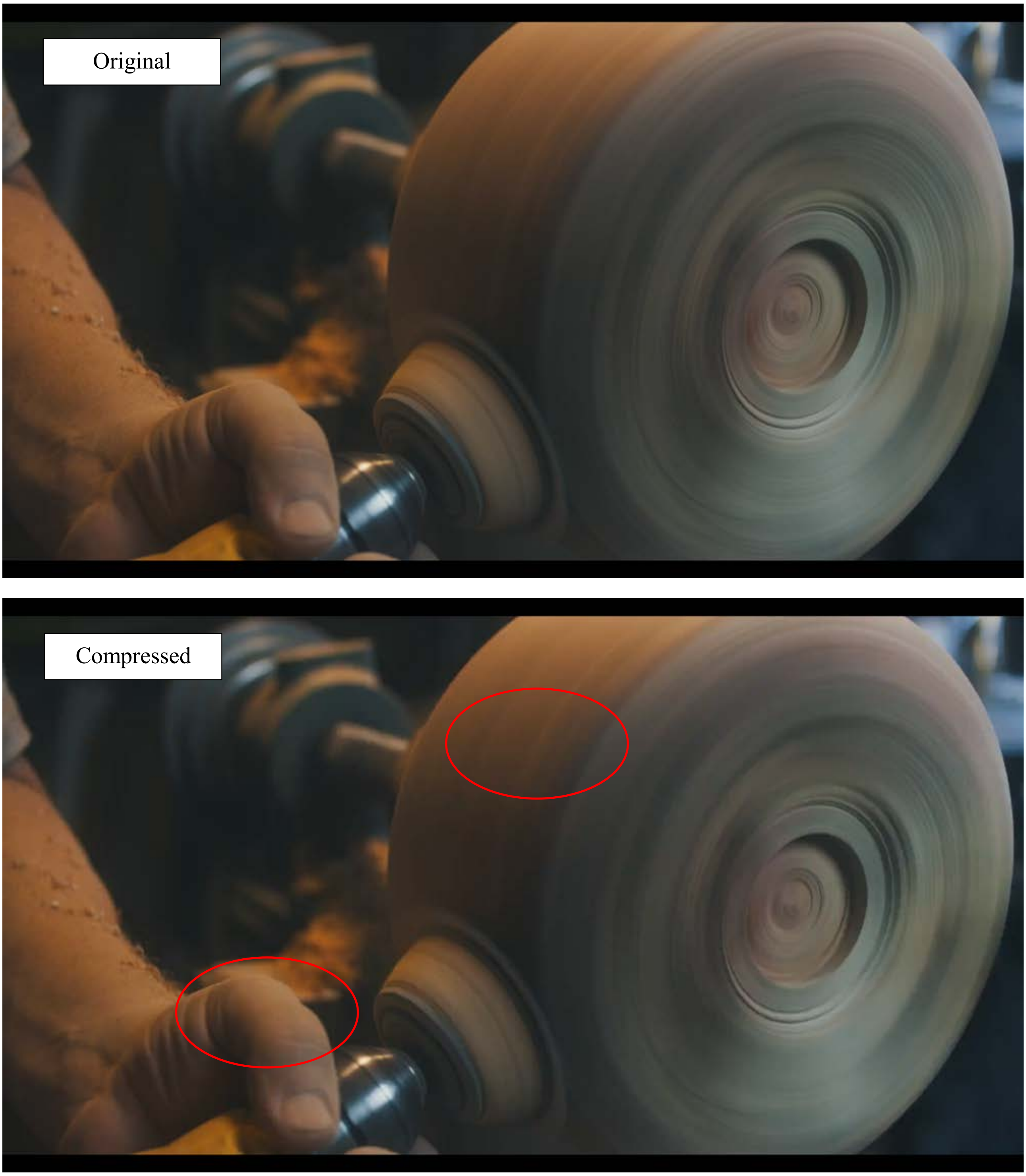


Figure 4. Example of blurring artefact

Mosquito

Mosquito noise becomes visible as the pattern of block-level prediction error changes a lot from one picture to the next. This type of noise is especially prominent in regions encoded with very small blocks (4x4, 8x8). Its visual effect is similar to random noise.

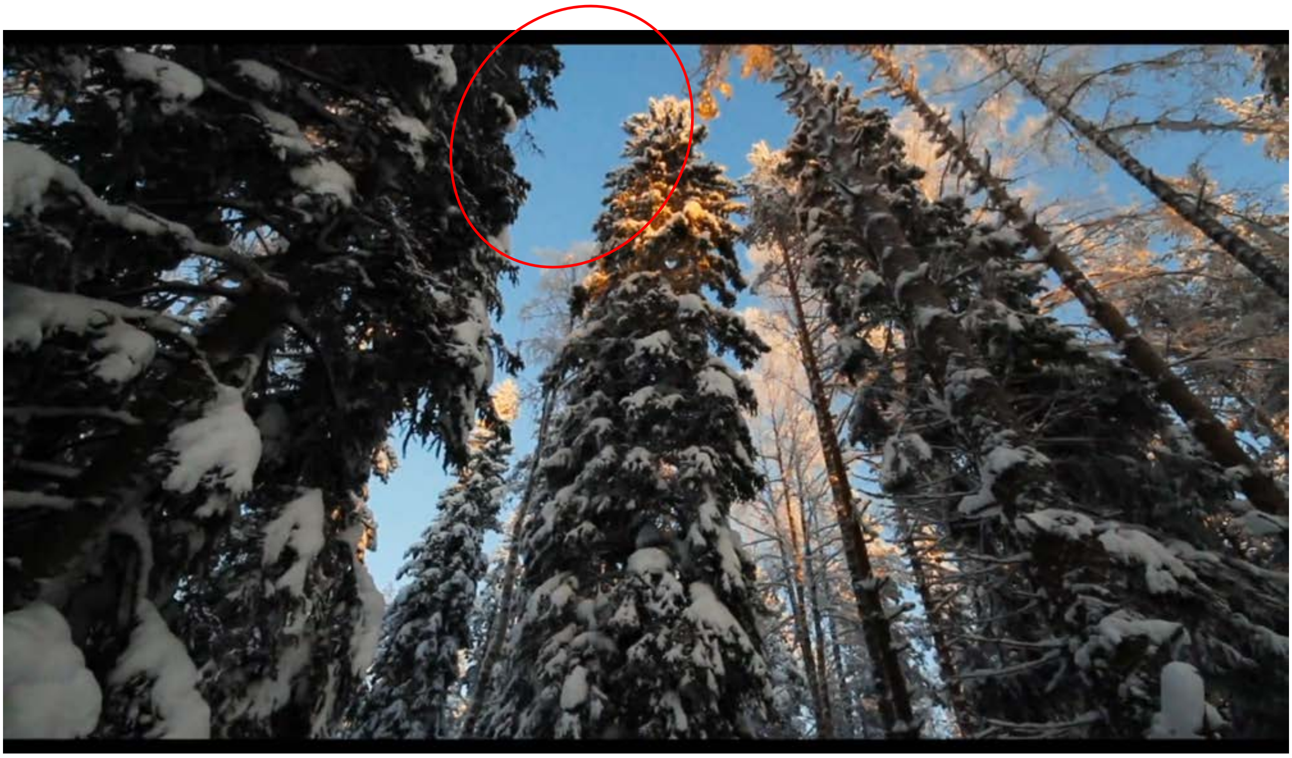


Figure 5. Example of mosquito artefact

Ringing

Ringing is another common artefact in video compression, which produces an effect of "halo" around sharp edges. More specifically, given that a steep edge contains a wide range of frequencies, the quantization of them, especially the high frequencies, involves an inaccurate reconstruction and, as a result, a transitory around the edge (Wien, 2015).

Figure 6. Example of ringing artefact

Stair case

This artefact refers to the inaccurate reconstruction of diagonal edges after quantization. More specifically, when performing quantization of a picture block, high frequency coefficients of the Discrete Cosine Transform (DCT) transform become zero, so the available horizontal and vertical basis functions of the inverse DCT are not able to recover the original pattern. As a result, not straight diagonal edges but stair-like edges becoming visually prominent as the quantization increases.

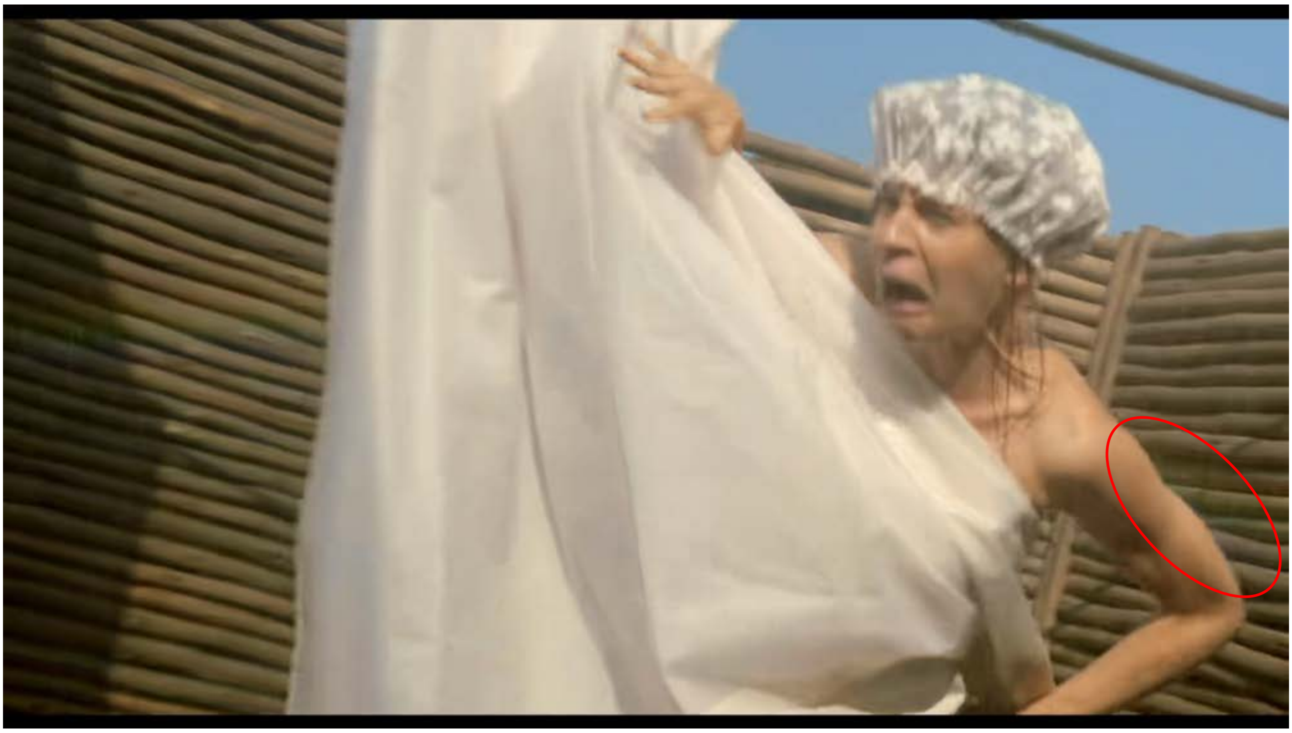


Figure 7. Example of stair case artefact

2.3.1.3 Transmission Artefacts (Buffering)

In VoD applications compressed movies are transmitted through the Internet for final display in web browsers. However, it might happen that some video packets are not delivered to the decoder on time, because of eventual network congestion or limited speed of the Internet connection. As a consequence, the decoder must wait (buffering) until receiving enough packets to continue the playback. Meanwhile, the user has to resign himself to watching the last decoded frame frozen.

2.3.1.4 Playback Artefacts (Skipped Frames)

If the computing resources of the receiving device are very limited and the decoder is not able to decode all the frames at the specified frame rate, the media player is, therefore, obligated to drop some frames. An intermittent video playback is then produced with a corresponding degradation of the playback quality.

2.3.2 Subjective Quality Assessment of Video Encoders

Objective quality metrics on compressed videos might not detect the exact subjective effect of certain compression artefacts. Although some objective quality metrics such as the Structural Similarity (SSIM) (Wang, 2004) correlate with the quality perceived by the human visual system better than the traditional Mean Squared Error (MSE), they are not as accurate as formal subjective tests based on the Mean Opinion Score (MOS) such as the ITU-R Rec. BT.500-13 (ITU-R, 2012) and the ITU-T Rec. P.910 (ITU-T, 2008).

2.3.2.1 Recommendations for Formal Subjective Tests

There are many ways for performing subjective tests, and many factors that can influence the final observer's decision, such as the distance between observer and monitor, the monitor settings (luminance, resolution, size), and the lighting conditions of the room. These factors lead to a high level of arbitrariness of the subjective test, whose results cannot be directly compared to others unless the testing conditions are exactly replicated. Therefore, the International Telecommunication Union (ITU) elaborated "formal" recommendations (standards) for the assessment of picture quality of a system under test under a controlled environment. BT.500 is tailored for broadcasting, whereas P.910 focuses on multimedia applications.

Generally, standardized testing methods describe the following aspects:

- Number and type of viewers: Although any number between 4 and 40 is possible, a minimum number of 15 observers is recommended to obtain meaningful averaged scores. These observers shall be non-experts ("naive"). In other words, they should not be familiar with, or not have deep knowledge on, i.e., picture quality evaluation or the system under study. However, a small group of experts (4-8) can participate in the early stages of the subjective experiment in order to configure the system under study and provide some indicative results.
- General viewing conditions: Viewing distance, peak luminance of the screen, resolution, observation angle, display brightness and contrast, chromaticity of background, and room illumination, among others.
- Instructions for the assessment: Before starting the experiment, the presentation structure of the test session should be clearly explained to the subjects. Generally, a training phase is first performed to stabilize the observer's opinion and to answer questions the viewers can ask about the experiment. Additionally, a description of the type of assessment and the grading scale must also be given. The experiment should not be longer than half an hour.
- Presentation of results: For each combination of the test variables, at least the mean value and the standard deviation of the statistical distribution of the assessment grade should be given.
- Selection of test material: The set of video sequences for subjective test is not standardized, but standards give some recommendations on the type of content according to the assessment problem. Some of the guidelines given in BT.500 are: 1) if the overall performance of a system is to be assessed, then "general and critical but not unduly so" content shall be selected; 2) if the experiment is to detect weaknesses of the system, then "critical and attribute-specific" material shall be used. P.910 recommends that the test sequences should cover a wide range of spatial and temporal information of interest to user of the device under test. However, in the Joint Call for Proposals on Video Compression Technology in 2010 (ITU-T, 2010) focused on defining a new video coding standard (that is, H.265), the set of video sequences were explicitly specified for subjective quality assessment based on BT.500.

- Selection of test methods: Subjective comparison methods can be divided into two groups: *single-stimulus* and *double-stimulus*. In single-stimulus the test sequences are presented one at a time and are rated independently on a category scale (i.e., excellent, good, fair, poor, and bad). No reference sequence is presented. In double-stimulus the reference sample (i.e. the uncompressed video) is presented as a part of the assessment process, and the test sample is rated keeping in mind the reference. This latter method is described in more detail in the following section.

2.3.2.2 Comparison Methods

In the context of subjective quality assessment of video codecs *double-stimulus* is the test procedure typically used (ITU-T, 2008), (ITU-T, 2010), (ITU-R, 2012), (J.R. Ohm, 2012), (P. Hanhart, 2012). In double-stimulus two methods are distinguished: Double Stimulus Impairment Scale (DSIS) and Double Stimulus Continuous Quality Scale (DSCQS).

DSIS

This method is used when the video content to be evaluated shows a range of visual quality that distributes well across the quality scale. The DSIS Basic Test Cell (BTC) of two consecutive presentations of a video clip (uncompressed and compressed) is illustrated in Figure 8: 1) a 1-s grey sequence with letter “A” is displayed; 2) a 10-s reference (uncompressed) sequence is presented; 3) a 1-s grey sequence with letter “B” is displayed; 4) the 10-s test (compressed) sequence is shown; and 5) the observed is asked to vote on the second sequence “B”.

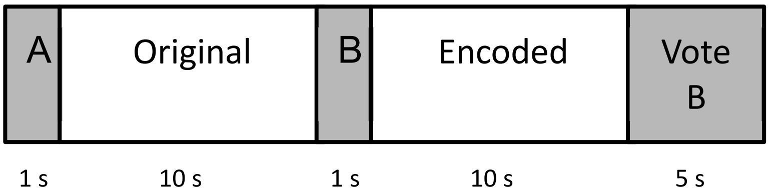


Figure 8. DSIS BTC

The quality rating scale consists of 11 levels, ranging from “0” (lowest quality) to “10” (highest quality).

DSCQS

This method is selected when the range of visual quality presented to the viewer correspond to upper part of the quality rating scale.

In DSCQS the original and encoder samples of a video clip are presented twice. The viewer is asked to evaluate not the encoded sample as in DSIS, but the two original and encoded samples separately. Moreover, the viewer does not know which one is the original or the encoded sequence. The position of the original and the encoded sequence is changed for each BTC.

As observed in Figure 9, a 1-s grey sample with letter “A” is displayed to announce the first sequence; afterwards, the 10-s original (or encoded) sequence is presented; again, a 1-s grey sample with letter “B” is displayed to announce the second sequence; after that, the encoded (or original) is reproduced; both sequences are repeated during a second pair of presentation by changing A and B into A* and B*, respectively; finally the observer is asked to vote on sequence "A/A*" and sequence "B/B*". This second repetition allows the viewer to gain the mental measure of the qualities associated with them.

In DSCQS the quality rating scale is made of two columns (one for A, one for B) with 100 horizontal marks.

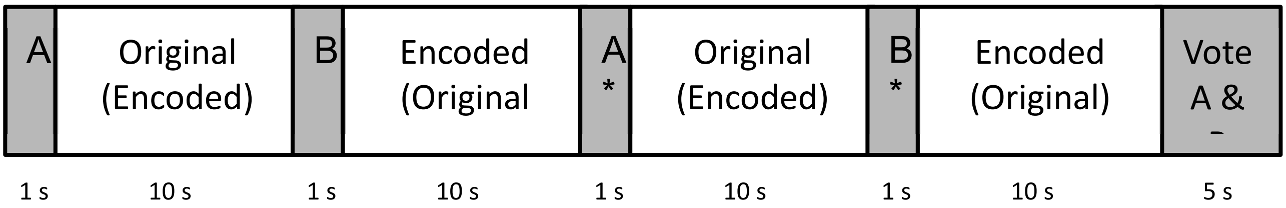


Figure 9. DSCQS BTC

2.3.3 Subjective Quality Assessment in the Context of Film265

In this section the informal subjective quality assessment of video codecs targeted for this project will be described in detail. The section is organized as follows: 1) the objectives of the informal subjective test in the context of the project are given; 2) the work methodology for this task is specified; and 3) each stage of the proposed work methodology is reported.

2.3.3.1 Objectives

The objectives of the informal subjective quality assessment in this project are:

- **To confirm that the results obtained in terms of objective quality assessment between video codecs are consistent with the visual quality:** About 50% bitrate reduction at equal objective quality is achieved by the TUB-H.265 encoder compared to the PicturePipe x264 encoder. It means that if a video sequence is encoded twice, one with TUB-H.265 at for example 2Mbps and another one with x264 at 4Mbps, the experienced visual quality should be identical (even better, since HEVC takes care more about the subjective quality (J.R. Ohm, 2012)). In order to verify this assumption, informal subjective tests from experts and non-experts have been conducted and reported in this document.
- **To provide a suitable configuration of the TUB-H.265 encoder that maximizes the visual quality for a given bitrate:** Based on the quality and bitrate settings defined by reelport for its VoD users, the TUB-H.265 encoder has been configured accordingly. In particular, two use cases for the upgraded PicturePipe platform are considered:
    - Using the same bitrate for both x264 and TUB-H.265, but increasing the quality (for example, from 720p to 1080p) in the latter encoder in order to improve the visual experience.
    - Using the same quality (resolution) for both x264 and TUB-H.265, but reducing the bitrate in the latter encoder in order to reduce costs.

2.3.3.2 Methodology

Based on the main objectives described above the working methodology was organized into five main lines:

1. **Design guidelines for the web-based subjective test:** Using as a start point the official recommendations for subjective quality assessment, in this phase the main objective of the web-based subjective test as well as the duration of the sequences, the expected number of expert and non-expert viewers, and the comparison method were defined.
2. **Initial selection of video sequences and bitrates:** In particular, pieces of representative 720p or 1080p movies encoded at typical bitrates for VoD were chosen.
3. **Informal subjective evaluation of video coding experts:** The objective of this evaluation was to give some **indicative results** on how much bitrate could be reduced with H.265 while keeping the same visual quality as that with H.264. The outcome of this stage has helped us to configure the final subjective test for regular viewers.
4. **Final selection of video sequences and bitrates:** The bitrates and qualities proposed in the third stage were verified by means of an informal subjective test for viewers who are not experts in video coding or video production. A final set of video sequences for the test was also selected.
5. **Informal subjective evaluation of non-expert viewers:** A web-based subjective test was developed for non-expert observers with the guidelines described in the first stage of the methodology. The outcome of this stage has been used as a feedback to adjust some encoding configuration.

2.3.3.3 Design Guidelines for the Web-Based Subjective Test

In the context of Film265, following the ITU recommendations for formal subjective test mentioned in Section 3.3.2 is out of the scope of the project for the following reasons:

1. These official recommendations require a lot of logistic effort, time and computing resources for collecting a representative number of test subjects and preparing test sessions in laboratory.
2. The subjective test projected for Film265 has to emulate a VoD environment. It means that the subjective test should consist of a web application where the user can watch the videos and make a comparison in an environment as close as possible as the final user environment: i.e. a web-based video player running on a desktop or laptop computer connected to the internet. Unlike a laboratory environment, the general viewing conditions of our target environment are so heterogeneous that they cannot be controlled.
3. In the context of Film265 not only the visual quality in terms of compression artefacts should be assessed, but also the influence of the available Internet bandwidth on the visual experience (e.g. buffering). According to this requirement, the original sequence cannot be presented in the subjective test because of its huge bitrate. Only video sequences compressed at typical bitrates for VoD shall be compared.
4. Unlike the ITU standards, questions from users cannot be answered “in-situ” before the test. Therefore, the subjective test should explain in words the methodology very clearly and be as simple as possible in order to avoid unsuccessful tests.
5. The web application should be very simple. In our particular case pairs of video sequences, one encoded with H.264 and the other one with H.265, with similar visual qualities have been presented to the observer. For each sequence the observer has been asked to 1) decide which version that has better quality, and 2) give a reason why. This latter question has been used as a feedback to find better configurations for PicturePipe.
6. Due to reasons “3” and “5”, the MOS, which uses absolute levels, is not the most appropriate

quality rating scale, but the following one based on differences: A is better, B is better, no difference.

Although the official recommendations for subjective quality assessment could not be strictly followed in the context of Film265, some of the decisions made for our web-based subjective test were inspired by them. In particular:

1. Number and type of viewers: Valid results could be obtained from at least 15 non-expert observers. Apart from this, in an early stage subjective tests have been performed by at least 4 experts belonging to the video coding field and film industry. From these comparisons, an appropriate configuration of the web-based subjective test came out for non-expert viewers.
2. Selection of test material: The set of video sequences must be representative for our particular assessment purpose, in particular: pieces of films with different spatio-temporal video properties.
3. Duration of test: According to the ITU recommendations, the experiment should not be longer than half an hour. We have created a web application whose expected duration was no longer than 10 minutes.
4. Comparison method: In our web-based subjective test, the two compressed samples of a sequence have been encoded with good quality, which is typical in the current VoD systems running under PicturePipe, so the range of visual quality was limited to a small portion of the quality scale. According to this, it seemed that DCSQS was the most appropriate comparison method, but with the following particularities:
    a. For practical reasons the original sequence was not assessed, but compressed versions.
    b. The application did not decide on when the viewer watches each sample, but the user had total control on the playback. Even the viewer could repeat each clip several times.
    c. The observer was asked to vote on sequence A and B not separately as recommended by DCSQS, but jointly using the scale based on differences mentioned before.
    d. The video clips were 20 seconds long instead of 10, so that the rate control could have enough time to stabilize the bitrate and the viewer enough time to detect compression artefacts.

2.3.3.4 Initial Video Dataset

The video sequences we have selected are extracts of movies and documentaries that exhibit different spatial and temporal video properties. In the following tables the datasheet of each movie is presented.

Title	My African Adventure
Director	Martin Miehe-Renard
Production	All Right Film A/S, Michael Obel
Country	Denmark
Language	Danish
Year	2013
Duration	0:01:41
Genre	Adventure, comedy

Selected by	LevelK
Input format	ProRes, 1920x1080p25, 4:2:2, 10-bit, YUV, 16:9
File size	2.3 GB
License	Proprietary
Available from	Copy from DCP hard disk, available TUB internal server
Type of content	Real-life footage, trailer with a lot of scene changes

Table 1. Datasheet of My African Adventure

Title	Maen & Hons (Men & Chickens)
Director	Anders Thomas Jensen
Production	M&M Productions, Kim Magnusson, Tivi Magnusson
Country	Denmark
Language	Danish
Year	2014/2015
Duration	0:02:51
Genre	Comedy
Selected by	LevelK
Input format	ProRes, 1920x1080p25, 4:2:2, 10-bit, YUV, 16:9
File size	2.8 GB
License	Proprietary
Available from	Copy from DCP hard disk, available TUB internal server
Type of content	Real-life footage, trailer with a lot of scene changes

Table 2. Datasheet of Men & Chickens

Title	Tale of a Forest
Director	Ville Suhonen, Kim Saarniluolo
Production	Matila Röhr Productions
Country	Finland
Language	Finnish
Year	2013
Duration	0:01:37
Genre	Documentary
Selected by	LevelK
Input format	ProRes, 1920x1080p25, 4:2:2, 10-bit, YUV, 16:9
File size	2.2 GB
License	Proprietary
Available from	Copy from DCP hard disk, available TUB internal server
Type of content	Real-life footage (nature), trailer with a lot of scene changes

Table 3. Datasheet of Tale of a Forest

Title	Tears of Steel
Director	Ian Hubert
Production	Blender Foundation, Ton Roosendaal
Country	The Netherlands
Language	English
Year	2012
Duration	0:12:14
Genre	Science fiction
Selected by	TUB
Input format	Raw TIFF, 4096x1714p24, 16-bit, RGB
File size	742 GB
License	Free movie under the Creative Commons Attribution 3.0 license
Available from	https://media.xiph.org/tearsofsteel/tearsofseel-4k
Type of content	Combines real-life footage with computer generated footage, short movie with moderate frequency of scene changes

Table 4. Datasheet of Tears of Steel

Title	Alan Turning Wood
Director	Phil Holland
Production	Red Digital Cinema
Country	USA
Language	English
Year	2014
Duration	0:04:54
Genre	Documentary
Selected by	Cinando
Input format	DPX, 4096x2160p24, 10-bit, RGB
File size	220 GB
License	Proprietary
Available from	Copy from DCP hard disk, available TUB internal server
Type of content	Real-life footage, short movie with moderate frequency of scene changes, very good quality and colours

Table 5. Datasheet of Alan Turning Wood

Title	The Lion
Director	Anthony Gilmore
Production	Nameless Media and Productions
Country	USA
Language	English

Year	2014
Duration	0:03:44
Genre	Documentary
Selected by	Cinando
Input format	DPX, 4096x2160p24, 10-bit, RGB
File size	178 GB
License	Proprietary
Available from	Copy from DCP hard disk, available TUB internal server
Type of content	Real-life footage, short movie with moderate frequency of scene changes, very good quality and colours

Table 6. Datasheet of The Lion

2.3.3.5 Informal Subjective Evaluation of Video Coding Experts

Experts in video coding field and film industry were asked to assess the visual quality of some video clips compressed with H.264 and H.265. A short biography of each asked expert is given below:

- **Frantz Delbecque:** Director of R&D and New Technologies at Eclair Group and Product Manager “Content Services” at Ymagis Group /dcinex. Eclair group is a French leading post production house and film laboratory.
- **Eric Cherioux:** Head of Post-production at CST (Commission Supérieure Technique de l'Image et du Son). Eric is responsible for the creation, design, and mastering of test equipment for digital cinema; conformation and internal calibration of films, DVD mastering and DCP (Digital Cinema Package). Founded in 1944 CST is a French organization for professionals in the audiovisual production.
- **Tilman Scheel:** Managing Director, founder, and owner of Reelport GmbH and Reelport France SAS. Reelport has a long history in handling films of film professional’s projects.
- **Dr. Mauricio Álvarez-Mesa:** Postdoc researcher at TU-Berlin, CEO and co-founder of Spin Digital Video Technologies GmbH. He has more than 10 years of experience in research about video codec implementation and he has published more than 20 papers and a book in this field.
- **Dr. Sergio Sanz-Rodríguez:** Postdoc researcher at TU-Berlin, Senior Researcher and co-founder of Spin Digital Video Technologies GmbH. Sergio has several years of research experience in video encoder design, mainly in rate control algorithms, he has several publications in this field.

Evaluation of Frantz Delbecque

He was first asked to compare the following sequences compressed with two encoders (x264, TUB-H.265) at different bitrates:

Sequence	Format	Bitrates for TUB-H.265 [kbps]	Bitrates for x264 [kbps]
My African Adventure	1080p24, 4:2:0, 8-bit	1200	2.5x, 2.0x, 1.5x, 1.0x
Men and Chickens	1080p24, 4:2:0, 8-bit	878	2.5x, 2.0x, 1.5x, 1.0x

Tale of a Forest	1080p24, 4:2:0, 8-bit	1290	2.5x, 2.0x, 1.5x, 1.0x

Table 7. Target bitrates for African Adventure, Men and Chickens, and Tale of a Forest

Below the Frantz's comments are given[1]:

It is clear that the ***H.265 version is better in terms of quality****.* ***I especially compared it with the H264 version that had the highest rate.***

The most salient points:

- *Flat tints and compression defects presented on the faces are not visible (or at least it does not mind the eye in the first reading).*
- *Tingling and compression defects on wide shots are largely mitigated (although still present ...) [in H.265]*
- *No significant difference on colorimetry.*
- *I would say that in terms of contrast, the H.265 version is slightly sweet but it's very subjective.*

By cons, trailers make the subjective visual analysis a little more complicated because the duration of the plans is too short.

Anyway, for a ***x2 gain compression****, we* ***[H.265]*** *have a* ***much better quality****, it is no doubt.*

3 examples on the one entitled "African Adventure" seems the most relevant. "Tale of a Forest" on this one as the H264 encoding defects were the most visible.

Frantz Delbecque was also asked to evaluate longer sequences. So, in a second round we prepared the following set:

Sequence	Format	Bitrates for TUB-H.265 [kbps]	Bitrates for x264 [kbps]
Tears of Steel	1080p24, 4:2:0, 8-bit	2000	4000 (2x), 5000 (2.5x), 6000 (3x)

Table 8. Target bitrates for Tears of Steel

And his comments are reported below:

I mostly did the comparison between the H.265 file at 2Mb/s and the H264 at 6mb/s and 5Mb/s. The ***H264 at 4Mb/s presents compression defects stronger than the H265****.*

In terms of overall quality, the H.265 is between the 2. Overall, the quality is identical to the definition / colour / contrast of skin and close-ups.

H.264 6 Mb/s has a little less compression defects on the backgrounds or solids.

---

[1] It should be noted that in his test he used a professional monitor

H.264 5 Mb/s is equivalent to H.265 on these points.

I no longer feel the "sweetness" that I had the last time of images of H265 compared to an image a little more contrasted of H264. But it's hard to evaluate these images with a lot of synthesis and VFX.

Evaluation of Eric Cherioux

Eric also compared the configurations illustrated in Table Table *7*. His comments about subjective quality are given next[2]:

First observation, the picture ***[H.265] is much smoother … digital noise is much less present, feel tingling is smaller and the picture looks static.***

The solid colours are more uniform. The codec gives the impression of a kind of post processing with a median blur (in LPF) that would protect areas of high contrast to keep intact the outlines of shapes.

The sharpness is still present, however, the result gives much smoother image than the other two H.264 …

This H.265 cleaning is done at the expense of details. If we observe for instance the skins of the characters, we see that we lose all asperities (for a luxury or cosmetic brand it would be probably better…). It is also obvious on the hair of animals that are dimmed.

Another finding, ***compression artefacts [H.265] are much less important but are much larger, therefore much less tingling but artefacts from areas with solid colour.***

I would base ***comparisons on the best video quality H.264:***

- ***H.265:*** *the images are* ***more pleasant in low light and appear sharper,*** *such as night scenes of "African adventure" (fewer artefacts).*
- *On the other hand, the lack of details in scenes with* ***animals in sunlight*** *in "Tale of a forest" would make me* ***prefer the H.264****.*
- ***Fast movements are also more pleasant in H.264;*** *the eyes capture more much bigger artifacts in H.265.*

If compared to the H.264 lower quality, the advantage is for the H.265.

The H.265 gives less sensation of "video" compressed for Internet.

In conclusion ***this is very subtle and visible only on very good screens or projections.***

I do not comment much on the weight of files that speak for themselves.

Evaluation of Tilman Scheel

About the configurations in Table Table *7. Target bitrates for African Adventure, Men and Chickens, and Tale of a*

---

[2] It should be noted that in his test he used a professional monitor

Forest, Tilman's comments are the following:

***I think 2.5x (H.264) is closest to H.265.** Take the artefacts in the black of second 28 in the "tale of a forest" as an example.*

Evaluation of Dr. Mauricio Álvarez-Mesa and Dr. Sergio Sanz-Rodríguez

A deeper analysis on a wider range of bitrates was conducted by the TUB team using a very good PC monitor on dark lighting conditions. The results are summarized in the following tables:

Bitrate [kbps]	Alan Turning Wood
1000	H.265: blocking, stair case, a lot of banding, blurring H.264: blocking, stair case, a lot of banding, blurring, mosquito, very bad quality
1500	H.265: less visible artefacts, better background H.264: blocking, mosquito, blurring, stair case, still very visible artefacts
2000	H.265: less blocking, more texture, still banding in background H.264: visible blocking, mosquito, blurring, stair case
2500	H.265: less blocking in motion areas, still banding in background H.264: mosquito, blocking, stair case
3000	H.265: more texture, still banding in background H.264: less mosquito, less blocking, still blurring
4400	H.265: more details, less banding H.264: more details, still mosquito
5000	H.265: more details, less banding H.264: more details, less mosquito

Table 9. TUB experts' informal subjective evaluation for Alan Turning Wood

Bitrate [kbps]	The Lion
1000	H.265: blocking, banding, stair case H.264: blocking, banding, mosquito, details are gone
1500	H.265: blocking, banding H.264: blocking, banding, mosquito, details are gone
2000	H.265: less blocking and banding H.264: less blocking, still mosquito, more details

2500	H.265: less banding H.264: still blocking and mosquito
3000	H.265: good quality, less banding H.264: still blocking and mosquito
4400	H.265: good quality, banding is some background areas H.264: more detail, still mosquito in background
5000	H.265: very good quality, still banding in some background areas although less H.264: more detail, still mosquito in background

Table 10. TUB experts' informal subjective evaluation for The Lion

Bitrate [kbps]	My African Adventure
1000	H.265: normal quality, minor artefacts H.264: highly visible blocking and mosquito
1500	H.265: normal quality, very similar to 1000 kbps H.264: very visible blocking and mosquito, but better than 1000 kbps
2000	H.265: better quality, less artefacts in challenging areas in terms of compression H.264: sill visible artefacts and mosquito
2500	H.265: better quality, few artefacts in challenging areas in terms of compression H.264: sill visible artefacts and mosquito
3000	H.265: good quality, very few artefacts H.264: still visible artefacts and mosquito
4400	H.265: good quality, very few artefacts H.264: visible artefacts in some parts and mosquito (but less)
5000	H.265: very good quality H.264: mosquito

Table 11. TUB experts' informal subjective evaluation for My African Adventure

Bitrate [kbps]	Tale of a Forest
1000	H.265: too much blocking H.264: inacceptable bad quality

1500	H.265: too much blocking H.264: inacceptable bad quality
2000	H.265: blocking, banding H.264: much more blocking and mosquito, loss of detail, bad quality
2500	H.265: blocking, banding H.264: much more blocking and mosquito, loss of detail, bad quality
3000	H.265: banding H.264: more blocking and mosquito
4400	H.265: intra-blocking H.264: still mosquito in blue sky
5000	H.265: some blocking H.264: mosquito and ringing

Table 12. TUB experts' informal subjective evaluation for Tale of a Forest

2.3.3.6 General Conclusion from the Experts' Comments

From the experts' comments we can conclude that with a 2.0x bitrate difference between H.264 (higher bitrate) and H.265 (lower bitrate), the latter one still produces better visual quality. With a 2.5x difference both encoders produce similar subjective quality in many aspects (Eric Cherioux still finds H.264 more pleasant for example in fast movements), although the most prominent compression artefact that each encoder produces has a different nature: mosquito (tingling) noise is very present in H.264 (not in H.265), whereas banding in uniform areas is more visible in H.265. With a 3.0x bitrate difference H.264 is slightly better.

Eric Cherioux also made an interesting observation: ***"The H.265 gives less sensation of "video" compressed for Internet"***. People are used to watching videos over the Internet, which are mainly encoded with H.264. So the regular H.264 artefacts (above all blocking and mosquito) are somehow associated with low-medium quality video over the Internet (e.g, YouTube quality). H.265, which has been proved to be better codec in terms of both objective and subjective quality, is able to change those artefacts for others that can be more pleasant even at up to 60% bitrate reduction (that is, 2.5 times more bitrate in H.264), giving less sensation of regular video over the Internet.

Although Tasks 3.1 and 3.2 on codec comparison conclude that the TUB-H.265 implementation is capable of producing in movies about 50% bitrate reduction for same objective quality compared to x264, the subjective analyses carried out by experts confirm that the coding tools of H.265 produce better subjective quality than those of H.264, as already reported in (J.R. Ohm, 2012).

2.3.3.7 Final Video Dataset and Bitrates

According to the subjective evaluation of the experts, the final set of 20-s video sequences as well as their bitrates for the assessed encoders that we propose for the informal web-based subjective test is summarized in the following table:

Sequence	Duration	Format	Bitrate for TUB-H.265 [kbps]	Bitrate for x264 [kbps]
The Lion	00:01:27 - 00:01:47	1080p24, 4:2:0, 8-bit	2000	4400 (2.2x)
Alan Turning Wood	00:01:57 - 00:02:17	1080p24, 4:2:0, 8-bit	2000	4400 (2.2x)
My African Adventure	00:00:00 - 00:00:20	1080p24, 4:2:0, 8-bit	2000	4400 (2.2x)
Tears of Steel	00:05:00 - 00:05:20	1080p24, 4:2:0, 8-bit	2000	4400 (2.2x)

Table 13. Video dataset and bitrates for the web-based subjective test

The 20-s video segments that we selected have representative spatial and temporal complexities, so that different types of compression artefacts can be detected (if can be). The proposed bitrates correspond to the regular (medium) video quality offered by reelport to its clients. As can be seen, the bitrate difference between both encodes that we propose is 2.2x, a little bit less than 2.5x considering that the optimal rate point for equal subjective quality is somewhere between 2.0x and 2.5x according to the expert's comments.

On the other hand, “Tale of a Forest” and “Men and Chicken” were dropped from the final set of movies for various reasons: “Tale of a Forest” is a very noisy sequence, and “Men and Chicken” is very easy to encode and, therefore, very difficult to detect compression artefacts even at low bitrates.

2.3.3.8 Informal Subjective Evaluation of Non-Expert Viewers

The URL of the web page for the subjective quality assessment designed for non-expert viewers is:

https://testing.picturepipe.net/micropages/film265-survey/

and is composed of an introduction page, 4 test pages (one per test video), and an end page.

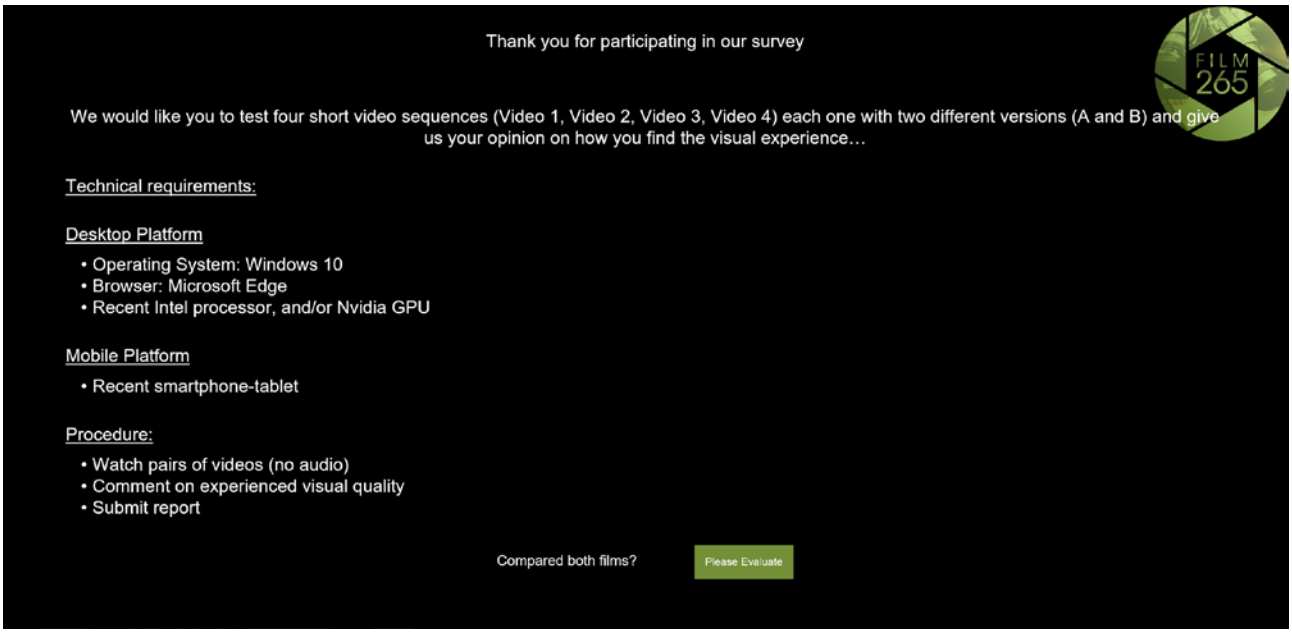
Introduction page

This page describes in detail the objective of the test, the technical requirements needed, and the procedure to be followed for the test.

- **Objective:** The viewers will be asked to test four short video sequences, each one with two different versions (A for H.265 encoding, and B for H.254 encoding), and give their opinion on how the visual experience is found. Details about the sequences and versions are described in Table 13. Versions A and B correspond to H.265 and H.264 encodings, respectively.
- **Technical requirements:** For desktop platforms Window 10, Microsoft Edge, and a recent Intel processor and/or Nvidia GPU are required to watch both versions of each sequence. For mobile platforms a recent smartphone-tablet are needed.
- **Procedure:** Pair of videos without audio will be watched, the experienced visual quality will be

commented, and finally a report will be submitted.

The figure below shows a screen capture of the introduction page. **Video 1 corresponds to “The Lion”, Video 2 to “Alan Turning Wood”, Video 3 to “My African Adventure”, and finally Video 4 to “Tears of Steel”**.



Thank you for participating in our survey

We would like you to test four short video sequences (Video 1, Video 2, Video 3, Video 4) each one with two different versions (A and B) and give us your opinion on how you find the visual experience…

Technical requirements:

Desktop Platform

- Operating System: Windows 10
- Browser: Microsoft Edge
- Recent Intel processor, and/or Nvidia GPU

Mobile Platform

- Recent smartphone-tablet

Procedure:

- Watch pairs of videos (no audio)
- Comment on experienced visual quality
- Submit report

Compared both films?

Figure 10. Introduction page of Film265-survey

Test page

The test page shows two video clips that the user will reproduce sequentially. After watching the two versions of the video sequence, the viewer will vote as follows:

Which one of the versions did you find the best?

- A
- B
- No difference
- Cannot play

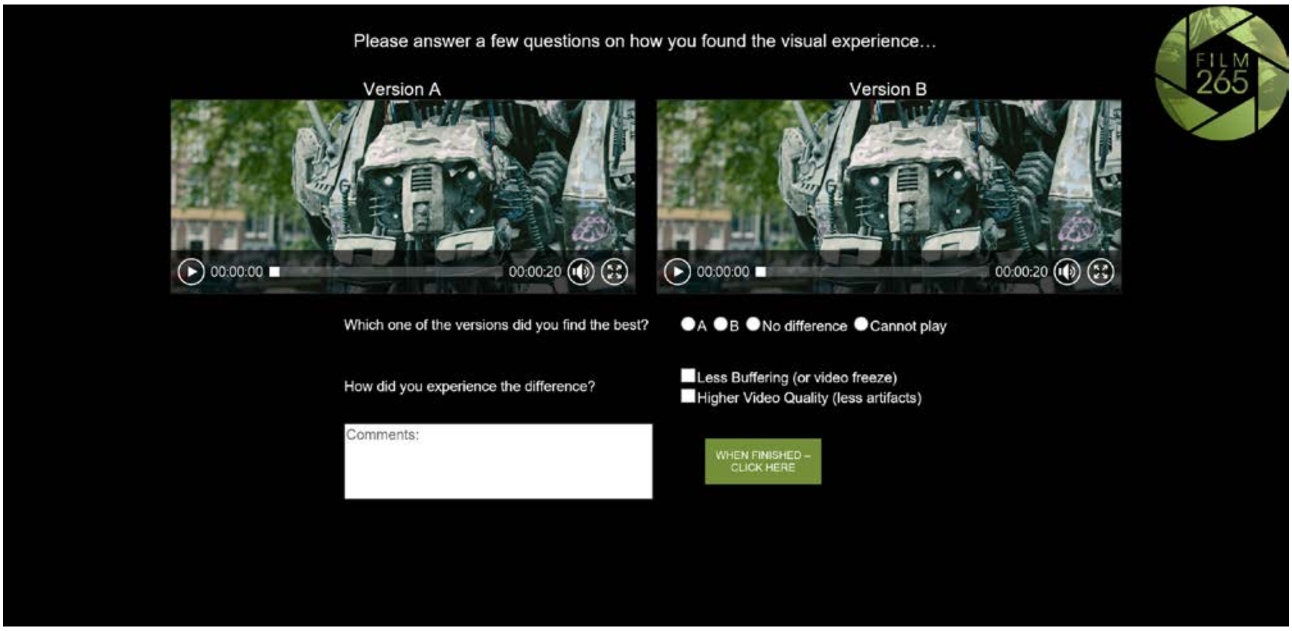
How did you experience the difference?

- Less buffering (or video freeze)
- Higher video quality (less artefacts)

Comments: ______________

For desktop platforms, if the viewer has a computer with the required technical specifications, he or she should be able to play the videos and mark options “A”, “B”, or “No difference”. In case the computer does not fulfil the requirements he should mark “Cannot play”. This last option has been also added to make this test page compatible for the analysis on multi-device support performed in Task 4.5. Although some recent devices such as smartphones, tablets and TVs can play H.265 videos, it was not possible to automatically detect this capability.

The figure below shows a screen capture of one of the test pages.



Please answer a few questions on how you found the visual experience…

Which one of the versions did you find the best? ○A ○B ○No difference ○Cannot play

How did you experience the difference?
☐ Less Buffering (or video freeze)
☐ Higher Video Quality (less artifacts)

Comments:

Figure 11. Test page of Film265-survey

End page

This page basically appreciates the viewer's help, and includes a link to the Film265 website. A screen capture of this page is shown below.

THANK YOU
– WE APPRECIATE YOUR HELP VERY MUCH

www.film265.eu

Figure 12. End page of Film265-survey

Evaluation of non-expert viewers

The web-based subjective test was performed by a total of 43 viewers, more than the maximum number, 40, recommended by BT.500. However, some of those tests had to be discarded, either since the PC did not fulfil the hardware requirements for successful H.265 playback or since some users did not click on the first question's options "A", "B", or "No difference".

The following table summarizes the results per video sequence and the average (total) results. Additionally, the mean and standard deviation on the opinion score were computed assuming label "1" for option "A" (H.265 is better), label "2" for option "No difference", and label "3" for option "B" (H.264 is better).

Video	No. Valid Tests	No. H.264 is better (Label 1)	No. No Difference (Label 2)	No. H.265 is better (Label 3)	Average	Standard Deviation
Video 1	36	7	19	10	2.08	0.68
Video 2	38	10	18	10	2.00	0.73
Video 3	35	7	18	10	2.09	0.69
Video 4	34	12	18	4	1.76	0.44
Total	143	36	73	34	1.98	0.70

Table 14. Average opinion score of the web-based subjective test

As can be observed, the average opinion score is 1.99, that is, almost label "2" that corresponds to opinion "No difference". Furthermore, more than the half of the total valid tests was scored as "No difference" (73 votes versus 36 for H.264 plus 34 for H.265). According to the results per video sequence, the average opinion scores for Video 1 "The Lion" and Video 3 "My African Adventure" are 2.08 and 2.09 respectively, so we can conclude that for these videos the subjective quality achieved by H.265 is equal or better than that by H.264. However, with an average opinion score of 1.73, the H.264 version of Video 4 "Tears of Steel" was voted more times than the H.265 version (12 versus 4).

In the following figures, we can see in percentages the results of the subjective test for every video as well as the total result.

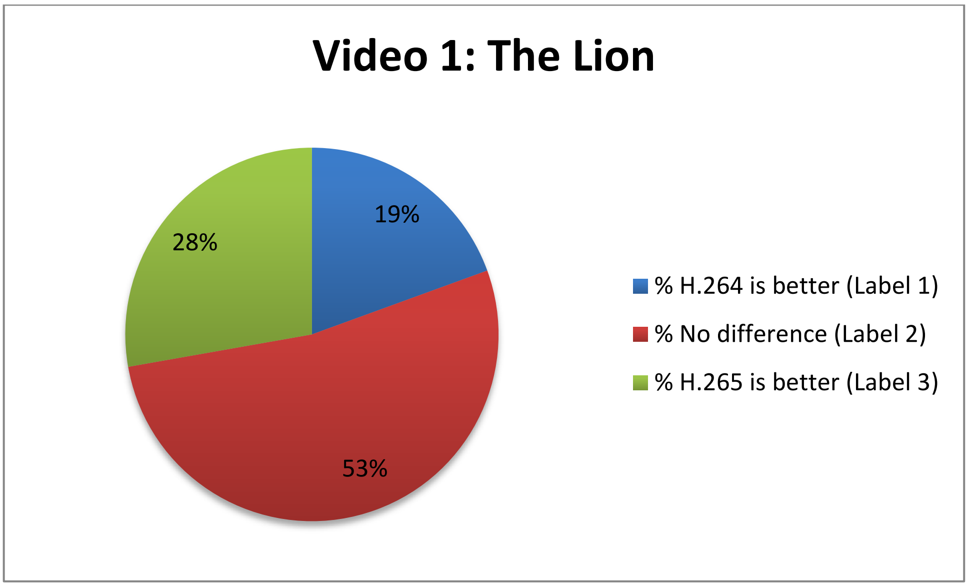


Figure 13. Result of the web-based subjective test for Video 1 “The Lion”

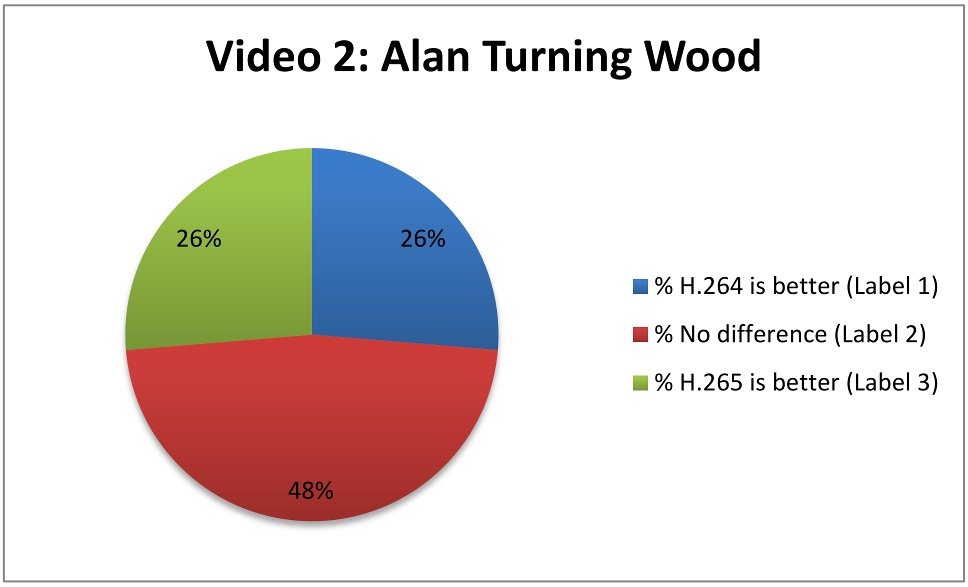


Figure 14. Result of the web-based subjective test for Video 2 “Alan Turning Wood”

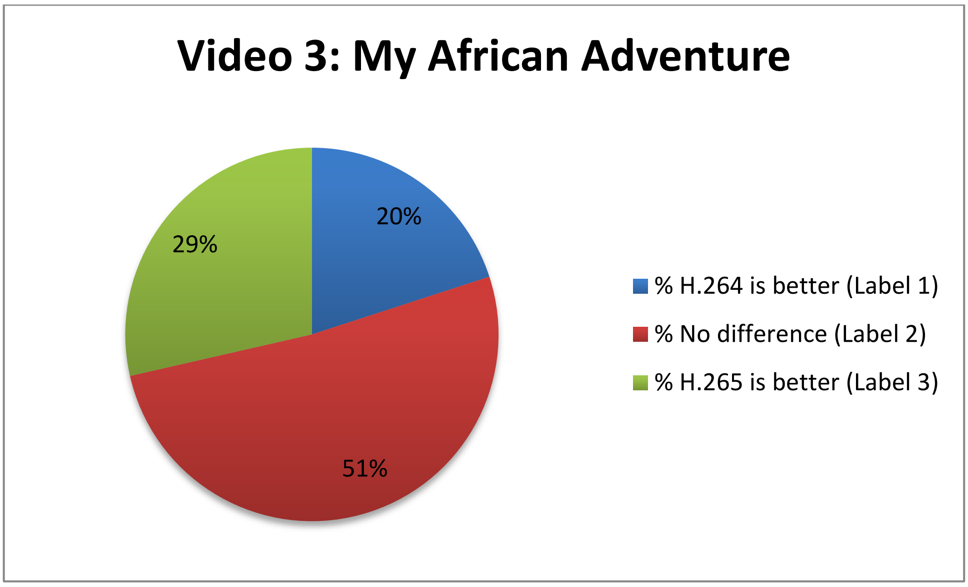


Figure 15. Result of the web-based subjective test for Video 3 “My African Adventure”

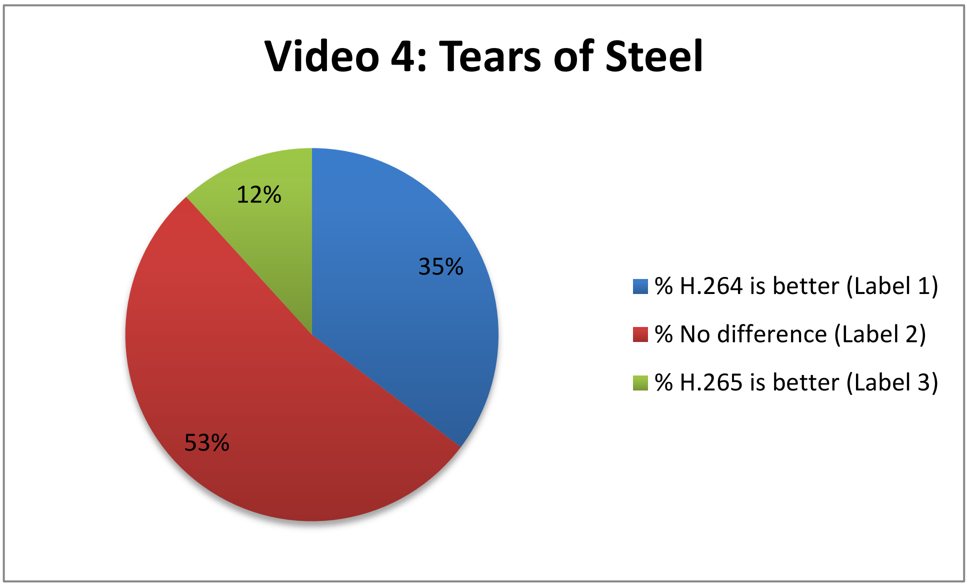


Figure 16. Result of the web-based subjective test for Video 4 “Tears of Steel”

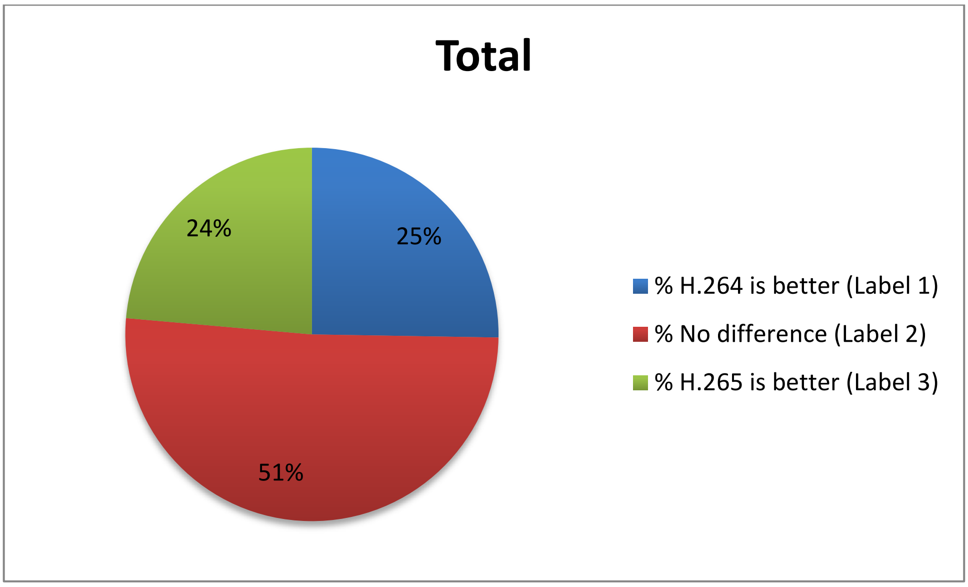


Figure 17. Average result of the web-based subjective test

As observed, more than 50% of votes per video clip went to “No difference”. On average, 25% of total users voted for H.264 and 24% for H.265, whereas 51% could not opt for one of the presented versions (see Figure Figure *17. Average result of the web-based subjective test*).

It is also worth noticing that, since in the subjective test version A corresponds to H.265 and version B to H.264 for every video, such an order of the versions might have biased the viewers’ evaluation. If it were true, part of the votes in favour of H.264 would have gone to H.265 and vice versa, whereas the percentage of “No difference” would have kept unaltered. In short, it is thought that the average results would have not changed much in case of a random order of the versions.

On the other hand, among the viewers who voted for H.264, 18% clicked on “Less Buffering”, 59% on “Higher Video Quality”, and 23% did not response. Regarding the users who voted for H.265, 31% marked “Less Buffering”, 66% “Higher Video Quality”, and only 3% did not give any response. The following figures show these percentages in a graphical way.

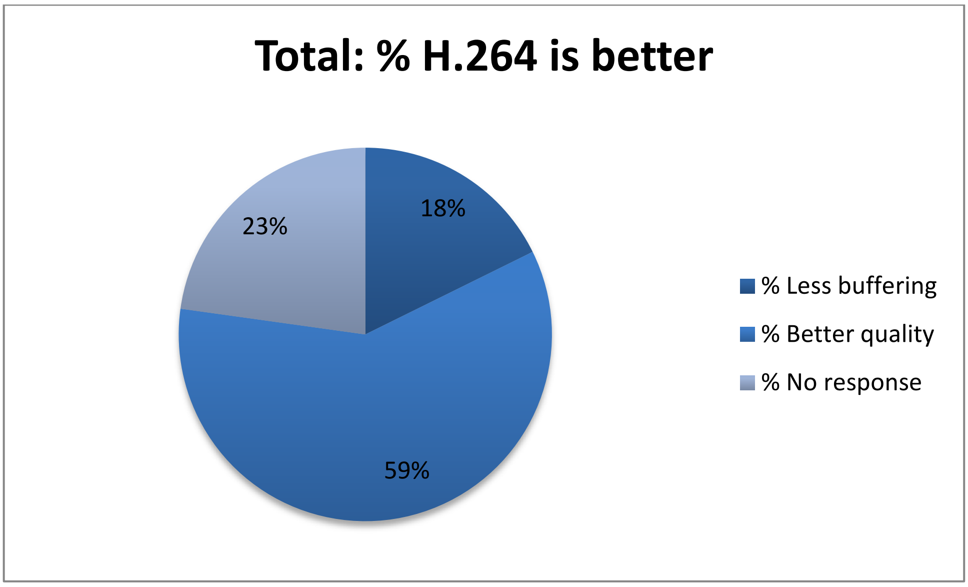


Figure 18. Result of the visual experience when the user voted for H.264

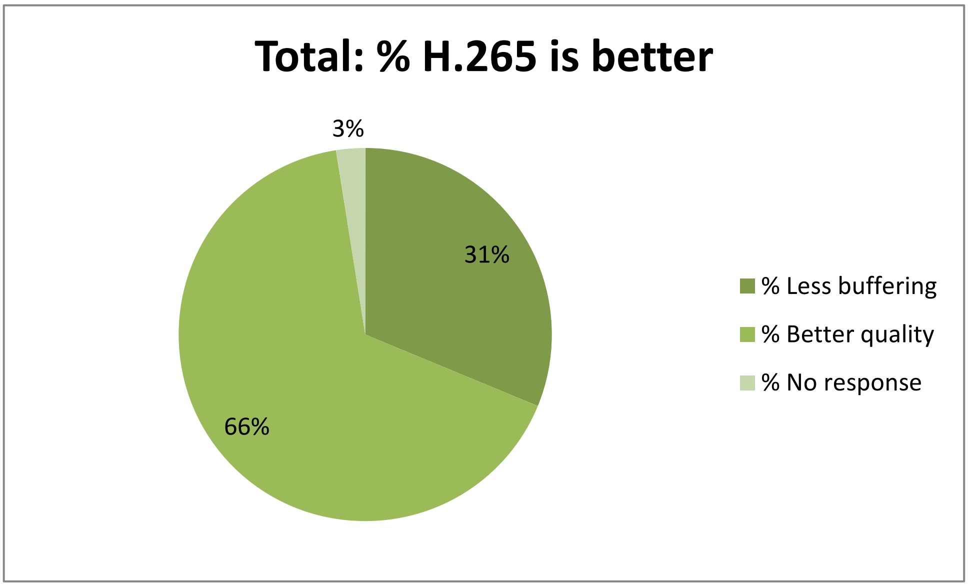


Figure 19. Result of the visual experience when the user voted for H.265

To finish with the results' analysis, only one viewer wrote a comment, specifically about Video 3 "My African Adventure":

"I would say that my assessment is subjective. I see no difference; it could be because I do not know how the differences look like. However, after replaying the videos couple of times, I would say that I still see no difference".

2.3.3.9 General Conclusion from the Non Experts' Comments

As already stated by the experts, with a 2.0x bitrate difference between H.264 and H.265 (lower bitrate), the TUB-H.265 encoder implementation still produces better subjective quality than x264, but a 2.5x difference does not ensures that both encoders give similar subjective quality. The results of the subjective tests conducted by the non-expert viewers prove however that the two assessed encoder implementations produce quite similar subjective qualities for a 2.2x bitrate difference.

To sum up, the upgraded H.265-based transcoding platform can guarantee equal or better user experience with bitrate ratios between 2.0x and 2.2x.

2.4 Advances over the State of the Art

Several studies have already assessed the performance of HEVC/H.265 in terms of subjective quality. These studies have relied on the formal recommendation ITU-R BT.500 for the setup and methodology of the tests. The work described in (S.-H. Bae, 2013) assessed the H.265 performance for 4K sequences and two different testing configurations: 1) two different viewing distances between observer and monitor; and 2) two different colour formats, in particular: YUV4:2:0 and YUV4:4:4. The paper concludes that H.265 video coding standard can be used for UHD-4K video at a bitrate of 18 Mbps.

Other studies, such as (J.R. Ohm, 2012) and (P. Hanhart, 2012), have compared the compression performance achieved in H.265 respect to that in H.264. Ohm et al. used standard definition (SD) and HD sequences for comparison, concluding that the H.265 standard produces around 35% bitrate reduction for same objective quality (PSNR) compared to H.264 and 50% reduction for same subjective quality (MOS). Hanhart et al. compared both standards using 4K videos, and the results showed a 44.4% bitrate reduction for same PSNR and 65.5% reduction for same MOS. In conclusion, H.265 is a standard much more suitable for UHD video than H.264.

However, in these studies typical film material was not used but very short test sequences, and the subjective tests were performed under laboratory conditions following the BT.500 recommendation. In other words, the scope of abovementioned tests, which aimed for measuring the compression capabilities of the H.265 standard, is different to that of Film265, which focuses on comparing -although in an informal way- in which a comparison is performed for video codec implementations under VoD viewing environments.

3 Conclusions

According to the results reported in this document, two main conclusions are given:

- The visual quality of short video clips compressed with x264 and TUB-H.265 at different target bitrates has been analysed informally by experts in the video processing field and film production. The experts gave a valuable analysis for the project’s purpose, concluding that TUB-H.265 produces better visual quality than x264 at half rate. With a 2.5x difference both encoders produce very similar quality, but x264 is still better in fast movements.
- Using the BT.500 recommendation as a reference, a web-based informal subjective test has been created for “naive” observers. The goal of the test page was to show H.264 and H.265 encoded videos at similar subjective quality, so that the user could report the experienced visual quality. Unlike formal subjective tests, this test page aimed at reproducing a common VoD environment, in

which the general viewing conditions cannot be controlled. From the comments of non-expert viewers, we can conclude that with a 2.2x bitrate ratio x264 and TUB-H.265 produce in most cases identical visual experience.

4 References

A. Punchihewa, D. G. (2002). *Artefacts in Image and Video Systems: Classification and Mitigation.* Auckland, New Zealand: Proceedings of Image and Vision Computing New Zealand, IVCNZ.

Apple. (2015, May). *Best Practices for Creating and Deploying HTTP Live Streaming Media for the iPhone and iPad*. Retrieved from Technical Note TN2224: go2sm.com/netflixencode

ITU-R. (2012, Jan). Methodology for the Subjective Assessment of the Quality of Television Pcitures. *ITU-R Rec. Bt.500-12*.

ITU-T. (2008, April). Subjective Video Quality Assessment Methods for Multimedia Applications. *ITU-T Rec. P.910*.

ITU-T. (2010, January). Joint Call for Proposals on Video Compression Technology. *39th Video Coding Experts Group (VCEG) Meeting: Kyoto (Japan)*, pp. 17-22.

J.R. Ohm, G. J. (2012, December). Comparison of the Coding Efficiency of Video Coding Standards - Including High Efficiency Video Coding (HEVC). *IEEE Transactions on Circuits and Systems for Video Technology*, pp. 1669-1684.

P. Hanhart, M. R.-S. (2012, October). Subjective Quality Evaluation of the Upcoming HEVC Video Compression Standard. *SPIE Optical Engineering and Applications*.

S.-H. Bae, J. K. (2013, June). Assessments of Subjective Video Quality on HEVC-Encoded 4K-UHD Video for Beyond-HDTV Broadcasting Services. *IEE Transactions on Broadcasting*, pp. 209-222.

Sanz-Rodríguez, S., Chi, C. C., Alvarez-Mesa, M., & Scheel, T. (2015). Deliverable D3.1: Report on Coding Efficiency and Performance Characteristics of the Optimized H.265 Encoder. *Horizon 2020 Project: Film265*.

Unterweger, A. (2013). Compression Artifacts in modern video coding and state-of-the-art means of compensation. In R. A. Farrugia, *Multimedia Networking and Coding* (pp. 28-49). IGI Global.

Wang, Z. (2004). Image Quality Assessment: From Error Visibility to Structural Similarity. *IEEE Transactions on Image Processing*, pp. vol 13, no. 4, 600-612.

Wien, M. (2015). *High Efficiency Video Coding.* Springer.