Perceptual Quality of 4K-resolution video content compared to HD

Glenn Van Wallendael*, Paulien Coppens†, Tom Paridaens*, Niels Van Kets*,
Wendy Van den Broeck† and Peter Lambert*
*Ghent University - iMinds - Data Science Lab, Belgium, Email: firstname.lastname@ugent.be
†Free University Brussels - iMinds - SMIT, Belgium, Email: paulien.coppens@iminds.be, wvdbroec@vub.ac.be

***Abstract*—With the introduction of 4K UHD video and display resolution, questions arise on the perceptual differences between 4K UHD and upsampled HD video content. In this paper, a striped pair comparison has been performed on a diverse set of 4K UHD video sources. The goal was to subjectively assess the perceived sharpness of 4K UHD and downscaled/upscaled HD video. A striped pair comparison has been applied in order to make the test as straightforward as possible for a non-expert participant population. Under these conditions and over this set of sequences, on average, on 54.8% of the sequences (17 out of 31), 4K UHD resolution content could be identified as being sharper compared to its HD down and upsampled alternative. The probabilities in which 4K UHD could be differentiated from downscaled/upscaled HD range from 83.3% for the easiest to assess sequence down to 39.7% for the most difficult sequence. Although significance tests demonstrate there is a positive sharpness difference from camera quality 4K UHD content compared to the HD downscaled/upscaled variations, it is very content dependent and all circumstances have been chosen in favor of the 4K UHD representation. The results of this test can contribute to the research process of developing metrics indicating visibility of high resolution features within specific content.**

I. Introduction

With the race for ever increasing resolutions and megapixels, people start to wonder whether it is possible to observe further improvements in this quality dimension. The latest resolution increase from High Definition (HD or 1920x1080 pixels) towards 4K UHD resolution (3840x2160 pixels) raised this question a lot. It can easily be explained that the capacity of the camera does not necessarily reflect the visual appearance. For example, when filming a smooth background with a 4K UHD resolution camera, the actual video sequence does not contain high resolution features. Consequently, equal quality perception could be obtained at a far lower resolution. When filming in a low light environment, the content will be influenced by sensor noise such that actual high resolution features will be hard to detect. When moving sharp objects over the screen at a high velocity, the eyes could have difficulties identifying the high resolution features of high resolution contents. Finally, artistic intent could produce a large amount of depth of field producing blurry regions in which resolution increase does not bring additional quality. When trying to solve the question on sharpness difference between 4K UHD and downscaled/upscaled HD, it is key to first subjectively evaluate this difference. Therefore, in this paper, a test will be described in which this difference has been investigated.

When thinking about the applicability of these results in the future, this research can be used to create quality metrics indicating visibility of high-resolution features within specific content. Such quality metrics could help video delivery services in making a thought through trade-off between resolution, dynamic range, color gamut size, and frame rate when transmitting over a limited network bandwidth. The results from this paper can be directly used to select a set of sequences on which upgrades to 4K UHD systems can be evaluated. Such upgrades can be found in video codec design, in display evaluation, or in video distribution chains.

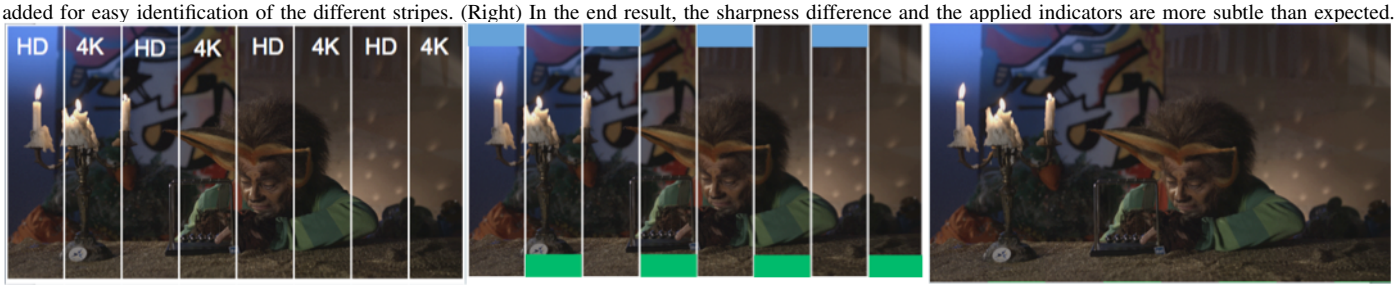
978-1-5090-0354-9/16/$31.00

II. State-of-the-art

Several years ago, when High Definition video services became the new broadcasting standard, H.264/AVC compression artefact visibility between HD and SD video content has been studied [1]. The SD content was created by scaling the HD content by a factor of two (both horizontally and vertically) and adding a grey border. As such, the SD videos were not upscaled during playback. Preliminary tests were conducted to select different encoding bitrates in order to reflect poor to good video quality. Video sequences were shown pairwise on 19 inch LCD using the SAMVIQ [2] methodology. This implies that test subjects could re-evaluate the videos as many times as wanted. Subjects were seated at a viewing distance of 3H or 3 times the height of the screen. In case of the native (not upscaled) SD content, this corresponds with a viewing distance of 6H, as per ITU-R Recommendation BT-500. After each comparison, subjects had to indicate their preference towards one of the two videos. Results show that preference towards HD or SD is influenced by the encoding artefact strength and image size. It was found that larger image sizes become a drawback in the case of encoding artefacts. In general, subjects prefer high quality SD to lower quality HD.

A similar study was also conducted in 2010 [3]. For the evaluation, the authors additionally considered the influence of packet loss on quality perception. In correspondence with the above described research results [1], the authors also conclude that bigger picture sizes result in higher perceived visual quality for high bitrates. Furthermore, the results show that the impact of the video format is to be neglected in the case of high packet loss rates. Finally, results analysis shows that there is a significant negative impact of video rescaling in the case of compression. Likewise, a later study investigated the joint impact of video resolution and upscaling operation on perceived visual quality for video resolutions up to SD (720x576 pixels) [4]. Again, their study showed that, in the case of coding degradations, upscaling always has a negative

Fig. 1: Striped pair comparison: (Left) The picture is divided in 8 stripes with alternating HD or 4K UHD resolution. (Middle) Blue and green indicators are added for easy identification of the different stripes. (Right) In the end result, the sharpness difference and the applied indicators are more subtle than expected.



impact on quality and that this impact further depends on the magnitude of the scaling.

A small test was conducted to investigate whether the difference between 4K and Full HD is visible on consumer grade (and consumer size) television sets when watching from a sensible viewing distance [5]. For the comparison, a native 55-inch HD television was positioned next to a native 55-inch 4K TV. As such, a 1-to-1 pixel match was made between the content and the native screen resolution. Participants were positioned at a viewing distance of 9 feet, corresponding to a viewing distance of 3.5H and were asked to pinpoint which of the two TVs was the 4K version. The same content was displayed on both TVs and downscaled for use on the HD television. Only one subject out of all 49 failed to correctly pinpoint the 4K TV.

In terms of video coding efficiency, a subjective quality evaluation has been performed between H.264/AVC and HEVC for resolutions beyond HD [6]. This study was based on a pair comparison (DSIS Variant II methodology) between different video sequences of five seconds long. Source content was stored in raw YUV 4:2:0 format, 8-bits per sample, with a frame rate of 24 and 30 fps. Each content type was encoded at five different target bitrates, determined using expert experiments. Results show that a significant bit rate reduction can be achieved when compressing resolutions beyond HD using the new HEVC video coding standard. For similar bitrates, test subjects observed significant differences between AVC and HEVC. Depending on the specific content type, bit rate reductions of up to 75% are reported by the authors.

An extensive subjective quality evaluation of HEVC has also been performed [7]. Several subjective experiments were conducted considering two viewing distances (0.75H and 1.5H), two colorspace formats (YUV 4:2:0 and YUV 4:4:4, both 8 bits per sample), three target bit rates (18, 23, and 36 Mbps), and nine different 4K-UHD test sequences. The sequences were played at 30 fps. The main results show that the viewing distance has only a marginal impact on quality perception and that the YUV 4:2:0 colorspace format does not result in a decrease of perceived visual quality. Based on the experiments, the authors conclude that the HEVC codec is able to deliver adequate 4K-UHD content under current broadcasting bandwidth conditions, even at a target bitrate of 18 Mbps. Also with respect to viewing distance, ITU-R Recommendation BT.1769 specifies 3H for HD (1920x1080), 1.5H for 4K UHD (3840x2160), 0.75H for 8K (7680x4320).

Finally, a study has been performed investigating different upscaling strategies for 720p and HD to 4K UHD conversion [8]. This test has been performed using a pair comparison methodology on a single screen using vertical striping of the content. This variation of pair comparison is according to our knowledge the easiest version to compare subtle differences like 4K UHD and HD quality differences and is therefore used in this paper as well, but further details will be provided later. The cited paper studies Preference of Experience which is relevant when comparing different upscaling and sharpening strategies. In this paper however, an upscaling strategy is used which is unlikely to produce sharpened results. Additionally, we believe it is difficult to ask for participant's experience score if the visualized image is artificially created by stripes of 4K UHD and HD resolution patches. Therefore, in this study, the subjects were asked for sharpness instead of preference. Additionally, sharpness can be considered a specialization of the main dimension for added-value of 4K UHD and is therefore considered first. The cited study concluded that in general conditions, low complexity upscaling algorithms like lanczos-3 had a higher performance compared to high complexity algorithms like super resolution algorithms. Eight sequences have been used for the comparison of the different upscalers. From these eight sources, it has only been proven for one source that 4K UHD provided a higher Preference of Experience than when upscaling from HD using the best performing upscaling algorithm. Therefore, our proposed research provides an addition to research from Li et al. by investigating a large set of publicly available sources. From our set of sequences, researchers investigating 4K UHD benefits would be able to make a source selection based on subjectively evaluated content sharpness.

III. Methodology

A. Pair comparison

According to our experience, performing a visual comparison between 4K UHD and downscaled/upscaled HD resolution is not straightforward, especially when the test subjects are expected to be non-experts. Therefore, to make this test as easy as possible to perform, a single screen pair comparison test methodology has been applied. Additionally, a pair comparison is considered the most reliable way to evaluate different quality scenarios. Other evaluation techniques have mainly been developed to reduce the test duration compared to a pair comparison. During a pair comparison, both the original and deteriorated sample are visualized. By having a look at both, the end-user is able to form an opinion and vote on one of both provided options. In this research, the paired comparison methodology has been simplified even more by a so-called striped variation [8].

In our striped pair comparison, first the 4K UHD sequence is downscaled to HD resolution using the Lanczos-3 [9] filter.

Fig. 2: Test environment and setup. Only one of the two monitors has been used at once for the striped pair comparison.

Lanczos-3 has been used because this upsampler had the highest overall preference for HD to 4K UHD conversion [8]. Next, sequence versions are generated by spatially interleaving downscaled/upscaled HD and 4K UHD stripes. Eight stripes are constructed like that. In the first version the odd vertical stripes are of a 4K UHD resolution and the second version has the native resolution in the even stripes. To help the observer even more, each stripe is marked with a blue or green bar at the top and the bottom like in the work of Li et al. [8]. During the evaluation process, it was asked which stripes are considered the sharpest: Blue or Green. An example of the creation process of such a striped pair comparison is given in Fig. 1.

B. Test setup

The test has been performed using the P.913 [10] method for the subjective assessment of video quality. To perform the test, a controlled environment has been arranged. The standard prescribes the room to be comfortable and quiet and this was fulfilled by performing the test in a dedicated room that was not used for any other purposes. The lighting level in the room can be considered dim and there was no significant noise present except the usual noise you can have in an office environment. According to the standard, preferably, an experiment should be designed such that each subject's participation is limited to 1.5 hours, of which no more than one hour is spent rating stimuli. The number of sequences has been restricted such that the test would stay within this constraint. In practice, we calculated that 60 sequences of 10 seconds each, repeated three times and followed by a voting period of 20 seconds at maximum would lead to 50 minutes for each participant. The standard also recommends to limit the number of times a given source stimulus is shown. In this experiment, every sequence is only rated twice, one time with 4K UHD resolution in the even stripes and one time with 4K UHD appearing in the odd stripes. In accordance with the ITU-R Recommendation BT.1769 [11], the viewing distance of 1.5H was respected.

As can be observed in Fig. 2, although the test would only require 1 monitor because of the striped pair comparison, two monitors have been used because of refresh rate restrictions. To perform a representative test, it was important that our raw video player would communicate to the monitor using Vertical Synchronization (vsync), such that the refresh rate of the screen was perfectly synchronized with the frame rate of the video. More specifically, the test has been performed on

Sequence	Info and source
bbb_scene3	Blender open movie projects Big Buck Bunny (3840x2160 60fps) Computer rendered, **furry textures**.
ElFuente_2 ElFuente_3 ElFuente_4 ElFuente_6 ElFuente_8	Netflix El Fuente [http://www.cdvl.org] (3840x2160 cropped from 4096x2160 59.94fps played at 60fps) *Artistic sequences with depth of focus (DOF) and moving cameras.*
InToTree ParkJoy	SVT - Fairytale (3840x2160 50fps) *Scanned from film, panning camera.*
Library Runners TrafficandBuilding TreeShade Fountains	SJTU 4K Video Sequences [http://medialab.sjtu.edu.cn/web4k/index.html] [12] (3840x2160 30fps) **No camera movement, no depth of field, highly textured regions like tree crowns, buildings, and bushes.**
Jockey YachtRide ReadySteadyGo	Tampere University of Technology - Ultra Video Group Test Sequences [http://ultravideo.cs.tut.fi/#testsequences] (3840x2160 30fps) *High motion (camera and objects).*
NTIA_Violin	NTIA - http://www.cdvl.org (3840x2160 cropped from 4096x2160 59.94fps played at 60fps) *Artistic DOF.*
News	Elemental - [http://www.elementaltechnologies.com/resources/4k-test-sequences] (3840x2160 30fps ProRes) Newsreaders in studio, **still camera, logo present**.
nebuta_festival steam_locomotive_train	MPEG test sequences (3840x2160 downscaled from 7680x4320 60fps) *Moving camera*, **tree crowns**.
UHD-1_Lupo_candlelight UHD-1_rain_fruits	EBU - UHD-1 test sequences (3840x2160 50fps) *Noise, moving objects,* **textured regions**.
clownlogoed susielogoed HarmonicMyanmar_Cobra Myanmar_boat Myanmar_bridge_bicycle Myanmar_bsm_f2_char Myanmar_child_dad Myanmar_tiger_waterfall SoccerSkills4	Demo Footage Center [http://www.harmonicinc.com/resources/videos/4k-video-clip-center] (3840x2160 30fps (first 3 sequences) 60fps (last 6 sequences)) *Artistic sequences*, **mostly still camera**.

TABLE I: Test sequences with source information. **Bold** tags provide an indication on sharpness features and *italic* tags provide counter examples.

Samsung U28D590D 4K monitors which in combination with our visualization PC was not able to show all required frame rates through one interface. More specifically, the HDMI2.0 interface could only be configured on frame rates higher than 30fps, while for 30 and 25fps, the DisplayPort interface needed to be used. Therefore, both identically calibrated displays were each connected through a different interface and by indicating to the user on which monitor the video could be played, there was, according to our observations, no impact on the test procedure. Calibration of the monitors has been performed using an X-Rite i1Display.

During the preliminary test phase of this experiment, the test has been performed on the Panasonic TX-65AX800E 4K TV screen (65 inch), which was able to play back all different frame rates using just a single interface cable. After some time however, the numerous switching between different refresh rates made the TV to stop responding to a 30fps request, such that the TV needed to be restarted. Therefore, for the actual test, the test setup has been changed to the set of monitors, such that participants would not be bothered with these technical restrictions.

C. Source content

To perform this test, 31 different sources have been chosen with the potential to range the entire resolution sharpness range (see Table. I). This distribution between easy to identify sources and difficult sources is necessary in order to get a representative test. A bias towards easy to identify sequences would turn the result overly positive in favor of 4K sharpness and the other way round. Only difficult to identify sequences would influence the general applicability of the conclusions. In fact, any conclusion could then be designed into the experiment confirming the intention of the designer. Additionally, a uniform sharpness distribution is needed in order not to discourage

the participants during the test. Too many difficult to identify sources would demotivate the participants resulting in more randomness in the results. It turns out that the proposed source content selection is rather uniform over the range of tested difficulty levels as will be proven in Section IV-C.

D. Participants

A total of 63 test subjects participated in the test. All test subjects were students at Ghent University and were between 20 and 25 years old (average age: 21.4). There were 59 male and 5 female participants. Using an Ishihara test, colorblindness of the participants has been evaluated. Visual acuity of the participants has been tested using a random set of small letters that had to be typed in a text box for evaluation. All 63 test subjects past the visual acuity test and the color vision test. As will be described later, although the age range and gender unbalance may not be representative for our world's population, the participants rating behavior shows a homogeneous spread across the whole range of possible values, making us conclude that this shortcoming may have a minimal impact on the result.

E. Test procedure

The test started with a registration of the participant in which their age and gender were asked. Next, the Ishihara and visual acuity tests have been performed. Participants were then provided with a description of the task they would be performing. The description also contained an illustration similar to the one presented in Fig. 1 such that the purpose and the objective of the test were very clear for the participants.

From this step on, a repetition of the following steps has been made such that all sequences would be covered. First, a random video sequence is played in a loop. The loop was performed at most three times or until the participant interrupted it using the ESC-button. The sequences were randomized for each participant. Second, participants were able to provide their opinion on which region was perceived as being the sharpest using two buttons on the screen mentioning "Green" and "Blue". Finally, the participant was requested to indicate whether this was a guess or not.

IV. Results

After performing this test, three aspects can be analyzed in detail, namely the content sharpness of HD downsampled video compared to the original 4K UHD resolution, test subject behavior throughout the test, and the test methodology.

A. Evaluation of content sharpness

In order to evaluate content sharpness, first some general observations will be provided followed by a detailed analysis of individual sequences and eventually also results from sets of sequences.

In general, in 62.6% of the votes, subjects observed that the 4K UHD parts of the video looked sharper than the downscaled/upscaled HD part. This observation implies that in the remaining 37.4% of the cases, the downscaled/upscaled HD regions have been identified as being sharper. A 1-sample test of proportions (Chi-square test) is able to reject the hypothesis of this 62.6% being smaller or equal to 50% (with a p-value smaller than 2.2e-16). So, there is a significant sharpness difference between the 4K UHD versions and the downscaled/upscaled HD versions of the sequences evaluated in this test when using the striped pair comparison technique with a three times repetition of the content.

Sequence	Score	Sign. >or <	p-value (>=0.5)	p-value (<=0.5)
bbb_scene3	0.714	>	1.000	0.000
ElFuente_2	0.603	>	0.987	0.013
ElFuente_3	0.524		0.672	0.328
ElFuente_4	0.548		0.836	0.164
ElFuente_6	0.738	>	1.000	0.000
ElFuente_8	0.492		0.465	0.536
InToTree	0.397	<	0.013	0.987
ParkJoy	0.659	>	1.000	0.000
Library	0.746	>	1.000	0.000
Runners	0.833	>	1.000	0.000
TrafficandBuilding	0.778	>	1.000	0.000
TreeShade	0.794	>	1.000	0.000
Fountains	0.833	>	1.000	0.000
Jockey	0.571		0.935	0.065
YachtRide	0.500		0.500	0.500
ReadySteadyGo	0.532		0.734	0.266
NTIA_Violin	0.548		0.836	0.164
News	0.706	>	1.000	0.000
nebuta_festival	0.500		0.500	0.500
steam_locomotive_train	0.548		0.836	0.164
UHD-1_Lupo_candlelight	0.722	>	1.000	0.000
UHD-1_rain_fruits	0.579	>	0.9547	0.045
clownlogoed	0.627	>	0.997	0.003
susielogoed	0.627	>	0.997	0.003
HarmonicMyanmar_Cobra	0.571		0.935	0.065
Myanmar_boat	0.587	>	0.969	0.031
Myanmar_bridge_bicycle	0.802	>	1.000	0.000
Myanmar_bsm_f2_char	0.556		0.877	0.124
Myanmar_child_dad	0.508		0.536	0.465
Myanmar_tiger_waterfall	0.532		0.734	0.267
SoccerSkills4	0.746	>	1.000	0.000

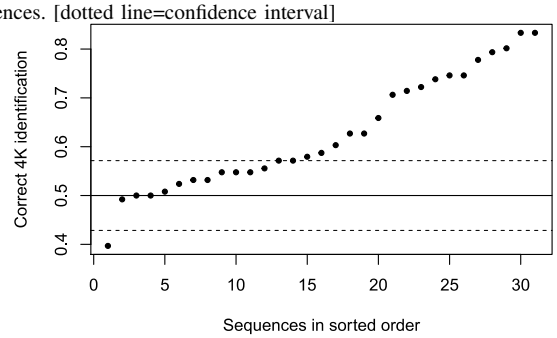
TABLE II: Test sequence score or 4K-UHD-identification correctness and indication of significance.

In theory, when performing this test of proportions on one fictitious average participant, the same significant conclusion could be drawn only because of the large amount of test sequences. However, in practice, when verifying this statement on the worst case population of this test, at least 32 participants would have been needed (p-value of 0.039). To do this check, our participants have been ordered from the one with the lowest ability to distinguish 4K UHD from downscaled/upscaled HD to the highest one. From this group, we needed at least the first 32 participants to come to the same conclusion.

When analyzing the individual sequences, it can be observed in Table II, that the InToTree sequence is the only sequence with a value significantly lower than 0.5 or 50% when performing a test of proportions. The significance of this value can be derived from the fact that the hypothesis of the score being larger than or equal to 0.5 can be rejected with a p-value of 0.013. In the entire column, the InToTree sequence is the only sequence with such a p-value smaller than 0.05. The same observation can be made from Fig. 3. In this figure, all sequences are ordered by score and visualized together with the calculated confidence interval. Observe that in this figure as well, the lowest scoring sequence (InToTree) is the only sequence below the confidence interval.

The other way round, in order to identify the sequences which have a significantly sharper 4K UHD representation than downscaled/upscaled HD representation, the last column in Table II needs to be analyzed. In this column, the hypothesis of having a sharpness identification less than or equal to 50% is rejected when the p-value in that column is less than 0.05. This is the case for all sequences indicated with a >-sign in the third column of this table. This can again be verified by

Fig. 3: Distribution of sequence 4K-UHD-identification correctness of different sequences. [dotted line=confidence interval]



looking at Fig. 3, where these 17 significantly higher rated sequences appear above the significance interval. Because 17 sequences have been rated with a higher sharpness, it can be concluded that 54.8% of the sequences is significantly sharper in 4K UHD than in downscaled/upscaled HD.

This leaves us with 13 sequences from which no difference between the 4K UHD and the downscaled/upscaled HD representation could be observed, even with the striped pair comparison. From our observations, sharpness features helping 4K UHD identification have been enumerated in Table I in bold. Features making it more difficult to see a difference between 4K UHD and downscaled/upscaled HD are given in italic.

It could also be observed that among the 10 sequences that were the most difficult to assess, the test subjects indicated to have guessed more than among the 10 easiest sequences, which is as expected. Among the 10 most difficult sequences, there is an average per sequence of 32.5% of the respondents who guessed. Among the 10 easiest sequences, the average is 21.4%. However, an odd finding is that for the most difficult sequence, namely InToTree, only 7 people wrote down they had to make a guess. It must be noted that the InToTree sequence, together with the ParkJoy sequence, originate from a 65mm film being digitized afterwards. Especially the InToTree sequence contains a lot of noise such that the participants could become confused about the concept of sharpness.

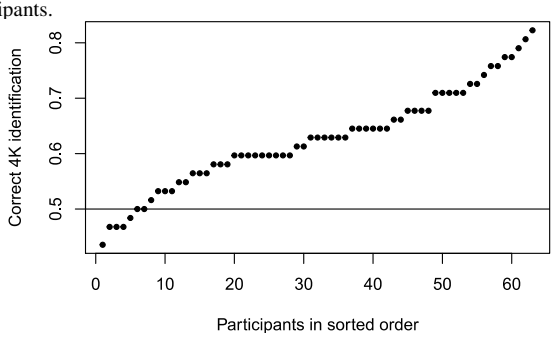
B. Analysis of the test subjects

The test subject capable of identifying the most 4K UHD representations was able to identify 82.3% of the sequences. The respondent with the lowest score correctly assessed 43.5% of the sequences. In total, five participants had a score lower than 50%. The average score is 62.6%. In Fig. 4, the scores of the different participants is visualized in a sorted order. Notice that the previously mentioned statistics can also be observed in this graph. Additionally, it can be seen that the spread of the scoring behavior is rather homogeneous. This observation strengthens our belief that the biased characteristics of our test subjects (young, mainly male) still provides generally applicable results, because we assume a larger population to provide homogeneous results as well.

By measuring the time it takes for a participant to complete the test, we observed that the fastest respondent finished the test in 28.8 minutes. The slowest respondent finished it in 42.1 minutes. The average test duration is 36.6 minutes, well within the estimated 50 minute duration.

An extra analysis to check for a correlation between the individual test duration and the overall test score was performed, but no significant correlation was found between the total test duration and the percentage of correct answers.

Fig. 4: Distribution of sequence 4K-UHD-identification correctness of different participants.



C. Evaluation of methodology

In order to evaluate the appropriateness of the test design, some observations will be made on the balance of sequence selection, training behavior, fatigue, balance between even and odd striped configurations, and balance in green/blue voting behavior.

To identify that the test sequences have been selected such that a good mix between easy and difficult sources has been made, the spread of the 4K UHD identification scores should be analyzed. This is important, because when there is non-uniformity present, participants have the tendency to lose motivation. In Fig. 3, this spread is visualized. It can be observed that the sequences almost uniformly cover the entire range without significant bias towards difficult or easy to identify sequences. It can be concluded that the test was thus balanced enough to not influence the scoring behavior.

To check for a possible effect of visual fatigue throughout the complete test, the number of correct answers during the first part of the test (the first 31 viewed sequences) with those of the second part of the test (the last 31 viewed sequences) have been compared by means of a test of proportions. This analysis reveals that test subjects don't vote similar for the first part and the second part of the test (p-value of 0.0403). 4K UHD sharpness has been identified more in the second part of the test. More specifically, the mean score for the first 31 sequences is 61.0%, whereas the mean score for the last 31 sequences is 64.2%. A possible explanation for the higher mean score in the second part of the test is that there might be some form of learning effect.

The learning effect has been further investigated on the aspect of repetition. In fact, every source is displayed twice, one time with 4K UHD sharpness in the odd stripes and once having this sharpness in the even stripes in random order. Consequently, every participant gets a second chance to identify differences in every source sequence. To do this analysis, all the scores of the first occurrences of an even or odd version have been grouped together. The second group contains scores of the opposite version which is a source repetition for the participant. When comparing these two groups, exactly the same voting behavior has been identified as when the votes were temporally separated as investigated in the previous paragraph. Consequently, this indicates that the learning effect can be ascribed to the fact that people observe the same source twice and improve their resolution sharpness detection skills on that specific source.

In general, one would expect that with a large enough population, there should be a similar scoring behavior between sequences that have a high sharpness in the even and in the odd striped versions. Overall, for the 31 odd striped sequences, on average, 65.8% of the sequences have been identified correctly, whereas for the 31 even striped versions, 59.5% is assessed correctly. Performing a significance analysis reveals there is a significant difference between these two results. When investigating individual sources, only for the sequences ParkJoy (odd: 77.8% correctly identified, even: 54.0%), UHD-1_Lupo_candlelight (odd: 84.1%, even: 60.3%), and clownlogoed (odd: 74.6%, even: 50.8%) there is a significant different score between the even and odd striped version. At least two possible explanations for such a behavior can be identified. First, there could be content specific characteristics which influence participants to consider certain regions as being the sharpest irrespective of the resolution difference. For example, in the UHD-1_Lupo_candlelight and the clownlogoed sequence, there is a highly textured tablecloth in a certain stripe of the generated sequences. Therefore, participants would consider the same stripe as being the sharpest in both the odd and even version. Second, the order of the green and the blue button used for the voting procedure have not been randomized with respect to their position. Consequently, guessing participants who have the tendency to click the same button could end up influencing the score in this way.

When investigating the blue/green voting behavior, in general, there is a significant bias in the 53.1% chance of a blue vote. Looking more closely at the individual sequences, no sequence has been identified with a biased vote for the blue option. Therefore, it can be considered more likely that blue votes were more popular when participants needed to guess.

V. Discussion and conclusion

To conclude, in this paper, in order to perform the difficult task of 4K UHD against downscaled/upscaled HD resolution comparison, a subjective test using a striped pair comparison methodology has been performed. In general, there is a significant sharpness difference between the 4K UHD versions and the downscaled/upscaled HD versions of the sequences evaluated in this test when using the striped pair comparison technique with a three times repetition of the content.

Overall, although there is a minor bias towards voting the blue option and there is a slight learning effect caused by displaying every source twice (odd and even striped), it can be concluded that both the sequence selection and the participant behavior results in uniform voting behavior. In order to cancel out the mentioned minor shortcomings, voting button position randomization could be considered and repetition of odd and even stripes could be avoided. Voting button position randomization is not expected to have a significant impact on the end results, whereas removal of repetition is likely to decrease significance of 4K UHD over downscaled/upscaled HD sharpness detection. This because the slight learning effect from the repetition of sources would be avoided such that high resolution sharpness would remain detected less.

Although significance tests demonstrate there is a sharpness increase from camera quality 4K UHD content compared to the downscaled/upscaled HD variations, it is very content dependent and only proven for this test setup. Additionally, a lot of participants agreed that the test was rather difficult. They confirmed that they had to guess a lot, especially with the sequences showing no significant sharpness difference.

The results of this test can contribute to the research process of developing metrics able to identify the necessity of 4K UHD transmission over downscaled/upscaled HD transmission depending on content characteristics. Such metrics can be beneficial when in the future a trade-off needs to be made between resolution, dynamic range, color gamut size, and frame rate.

Acknowledgment

The research activities described in this paper were funded by Ghent University, iMinds, Flanders Innovation & Entrepreneurship (VLAIO), the Fund for Scientific Research Flanders (FWO-Flanders), and the European Union. The computational resources (Stevin Supercomputer Infrastructure) and services used in this work were provided by the VSC (Flemish Supercomputer Center), funded by Ghent University, the Hercules Foundation and the Flemish Government department EWI. The work is performed as part of the iMinds VFORCE (Video 4K Composition and Efficient streaming) project (under IWT grant agreement no. 130655).

References

[1] S. Pechard, M. Carnec, P. L. Callet, and D. Barba, "From SD to HD Television: Effects of H.264 Distortions Versus Display Size on Quality of Experience," in *Image Processing, 2006 IEEE International Conference on*, Oct 2006, pp. 409–412.

[2] ITU-R WP6Q, "SAMVIQ - Subjective assessment methodology for video quality," *International Telecommunication Union*, Sept. 2003.

[3] M.-N. Garcia and A. Raake, "Quality Impact Of Video Format And Scaling In The Context Of IPTV," in *Third International Workshop on Perceptual Quality of Systems (PQS) 2010*. ISCA/DEGA, sep 2010, pp. 119–124.

[4] B. Belmudez and S. Moller, "An Approach for Modeling the Effects of Video Resolution and Size on the Perceived Visual Quality," in *Multimedia (ISM), 2011 IEEE International Symposium on*, Dec 2011, pp. 464–469.

[5] "4K Resolution Is Visible vs 1080p on 55 TV from 9 Viewing Distance," http://www.hdtvtest.co.uk/news/4k-resolution-201312153517.htm, accessed: 2016-03-03.

[6] P. Hanhart, M. Rerabek, F. De Simone, and T. Ebrahimi, "Subjective quality evaluation of the upcoming HEVC video compression standard," in *SPIE Optical Engineering+ Applications*. International Society for Optics and Photonics, 2012, pp. 84 990V–84 990V.

[7] S. H. Bae, J. Kim, M. Kim, S. Cho, and J. S. Choi, "Assessments of Subjective Video Quality on HEVC-Encoded 4K-UHD Video for Beyond-HDTV Broadcasting Services," *IEEE Transactions on Broadcasting*, vol. 59, no. 2, pp. 209–222, June 2013.

[8] J. Li, Y. Koudota, M. Barkowsky, H. Primon, and P. L. Callet, "Comparing upscaling algorithms from HD to Ultra HD by evaluating preference of experience," in *Quality of Multimedia Experience (QoMEX), 2014 Sixth International Workshop on*, Sept 2014, pp. 208–213.

[9] C. E. Duchon, "Lanczos filtering in one and two dimensions," *Journal of Applied Meteorology*, vol. 18, no. 8, pp. 1016–1022, 1979.

[10] ITU-T P.913, "Methods for the subjective assessment of video quality, audio quality and audiovisual quality of internet video and distribution quality television in any environment," *International Telecommunication Union*, Jan. 2014.

[11] ITU-R BT.1769, "Parameter values for an expanded hierarchy of lsdi image formats for production and international programme exchange," *International Telecommunication Union*, Jul. 2006.

[12] L. Song, X. Tang, W. Zhang, X. Yang, and P. Xia, "The sjtu 4k video sequence dataset," *Fifth International Workshop on Quality of Multimedia Experience (QoMEX2013)*, Klagenfurt, Austria, 2013.