Subjective Quality Evaluation of a Full Resolution Video with Eligible Number of Reference Pictures

Anna Ostaszewska-Liżewska
Warsaw University of Technology, Mechatronics Faculty, ul. św. Andrzeja Boboli 8, 02-525 Warszawa, Poland,

Rafał Kłoda
ŁUKASIEWICZ Research Network – Industrial Research Institute for Automation and Measurements PIAP, Al. Jerozolimskie 202, 02-486 Warszawa, Poland

Abstract: The paper presents a new concept of video presentation for subjective quality evaluation methods with a reference given in parallel. The idea is to split a full resolution video into an n-picture matrix and encode each cell differently to compare aspects of processing such as encoding parameters or lossy compression algorithms. Conducted experiments show that it is possible to get more accurate results, shortening the time of evaluation at least by half with a less complicated and cheaper experimental station.

Keywords: video compression, subjective quality evaluation

1. Introduction

Subjective user-perceived video quality ratings provided by a human audience and expressed as a mean opinion score (MOS) are still considered to be the most reliable measure of compressed video quality [1, 2]. Aiming at unification, in its recommendations the International Telecommunication Union (ITU) describes the methods of conducting subjective experiments. There are methods that use single stimuli and others which use more than one picture to score quality. Additional pictures are used as a reference, which enables comparison and helps to formulate assessment on a given scale. When stimuli are displayed sequentially, such as in the case of pair comparison (PC) [3], degradation category rating (DCR) [3], or subjective assessment methodology for video quality (SAMVIQ) [4], the experiment takes more time and the measurement accuracy is decreased by memory effects [7]. This problem does not occur in the case of methods that use stimuli given in parallel, such as in the case of the simultaneous double stimulus for continuous evaluation (SDSCE) [5] or triple stimulus continuous evaluation scale (TSCES) [6], but in turn they require duplication of video encoders and monitors, which multiplies the price of the experimental station. This is a big drawback, especially in the case of new technologies such as 8K ultra-high-definition (UHD) television, for example. The other problem is synchronizing video signals and the unnatural viewing angle, for example, when the observer has to compare 8K UHD pictures. For those reasons, the number of pictures is limited to two for SDSCE or three for TSCES. This paper proposes a new method with multiple stimuli displayed in parallel on the same independently processed. The screen-splitting matrix defines the number and surface area of sub-images simultaneously given to the observer. The final image does not contain any lines separating the divisions (Fig. 1).

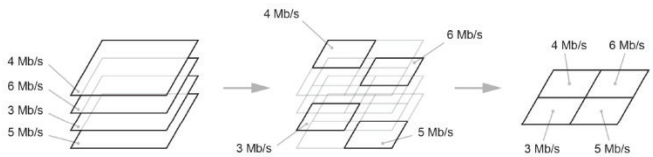

Fig. 1. Diagram illustrating the method of splitting into an *n*-parts (*n* = 4)
Rys. 1. Schemat ilustrujący metodę podziału ekranu na *n*-części (*n* = 4)

The number n of stimuli can be changed according to experiments needs. The limit is the observers' ability to perceive and evaluate more than one video at the same time.

2. Experiment

To evaluate the proposed method, three experiments were conducted: PC, PC-1×2, and PC-2×2. The PC experiment was a pair comparison (PC) method in accordance with ITU recommendations [3]. The observers were to choose the video of a better quality in each pair of images presented alternately one after each other. In the PC-1×2 experiment, each half of the frame was coded with different parameters and observers had to choose the better one. In the case of PC-2×2, the screen was divided into two rows and two columns and observers were asked to point out which quarter had the best quality and which had the worst. The purpose of the PC experiment was to obtain reference results for PC-1×2. The result of the PC-2×2 experiment was supposed to answer the question of whether the observer was able to recognize the quality of an image divided into quarters, each of which was coded with different parameters. Thirty-nine observers participated in the PC experiment, 16 in the PC-1×2 test, and 39 in the PC-2×2 test.

To allow a comparison of the obtained results, the same video was used in all experiments. This was a rendered animation composed of two characteristic scenes (Fig. 2). For both scenes, the

Autor korespondujący:
Rafał Kłoda, rkloda@piap.pl

Artykuł recenzowany
nadesłany 04.02.2019 r., przyjęty do druku 14.03.2019 r.

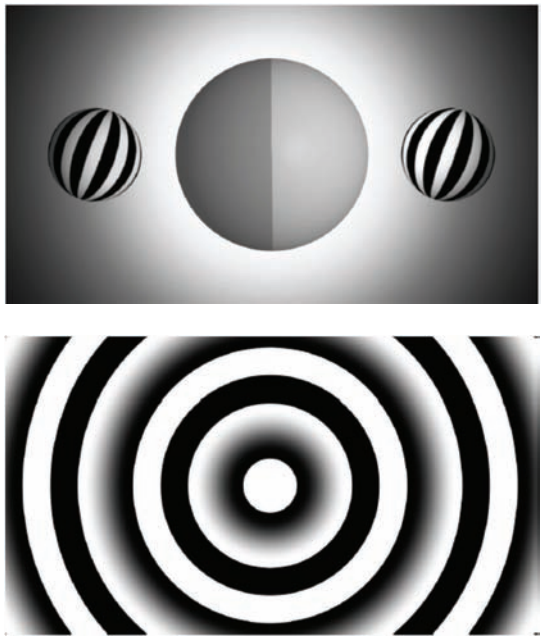

Fig. 2. Frames of video test material: a) scene 1, b) scene 2
Rys. 2. Ramki materiału testowego: a) scena 1, b) scena 2

image was symmetrical to the centre of the screen to ensure the same probability of image distortion in each part. Due to the technical limitations of the laboratory equipment, a standard progressive MPEG-2 MP@HL (Main Profile @ High Level) with 1920 × 1080 px, 24 fps, and constant bitrate had to be chosen, but the concept is valid for any resolution and compression algorithm.

The content was designed especially to obtain compression artefacts that are typical for MPEG-2: blocking, blurring, ringing, staircase, basis pattern artefacts, and mosquito noise [8]. Variation of the quality of the test sequences was achieved only by changing the bitrate in the coding process. Bitrate values were selected experimentally in pre-tests to provide clearly perceptible differences in quality. Four bitrate levels ($n = 4$) were chosen: 3, 4, 5, and 6 Mbps. Observers were not shown the source (uncompressed) video and were not informed about the compression parameters used.

In the case of PC and PC-1×2, in accordance with recommendations [3], all $n(n-1) = 12$ possible combinations were tested (Fig. 3). Because both sequences were displayed in parallel on the same screen in PC-1×2, the test time was halved in comparison with PC. In PC-2×2, the number of sequences evaluated simultaneously increased to four, therefore an observer would have to watch the same test sequence (with different sets of bitrates in individual quarters) 24 times. To lessen the observers' fatigue and to shorten the duration of the experiment, 10 variations were chosen randomly (Fig. 4). In this test material, each sub-image encoded with a certain bitrate was displayed at least once in any quarter of the screen

3. Data analysis

In PC-2×2, the best part was always represented by a quarter encoded with 6 Mbps and the worst by one encoded with 3 Mbps. Therefore the difficulty of recognizing the quality of the image fragments in each set was constant. In the PC and PC-1×2 methods, the difficulty level varied for each set depending on the quality levels of the selected pair of images. Therefore, the results of the PC and PC-1×2 methods cannot be directly related to the results of PC-2×2.

As expected, in the case of both the PC and PC-1×2 methods, the least votes were assigned to sequences encoded with the lowest bitrate (3 Mbps) and the most to the videos encoded with the highest bitrate (6 Mbps). The number of votes cast increases evenly with the increase of the bitrate, which means that in most cases observers were able to recognize and indicate an image of better quality. The increase in the number of votes is greater for PC-1×2, where two stimuli were presented simultaneously. The video with the lowest bitrate was chosen as the better one in only 2.6% of presentations in PC-1×2 and in only 4.9% of cases in the PC method. Assuming that choosing a video with a higher bitrate is the correct answer, the percentage of correct votes was calculated and is presented in Table 1 and Table 2.

The correctness of PC-1×2 votes is higher than that achieved with the classical PC-1 method. Only in the pairs 3 & 5 Mbps and 5 & 6 Mbps there is a lower percentage of correct votes:
– 3 & 5 Mbps pair: 93.75% (PC-1×2) and 97.44% (PC-1),
– 5 & 6 Mbps pair: 81.25% (PC-1×2) and 82.05% (PC-1).

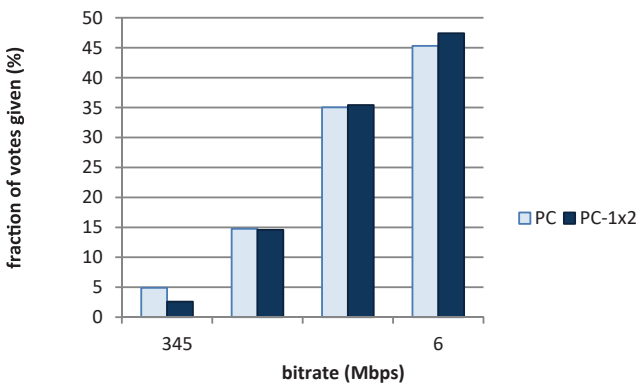

Fig. 5. Distribution of votes cast on the better quality image in the PC and PC-1×2 methods
Rys. 5. Rozkład głosów oddanych na sekwencję w eksperymentach PC i PC-1×2 w zależności od wielkości strumienia bitowego

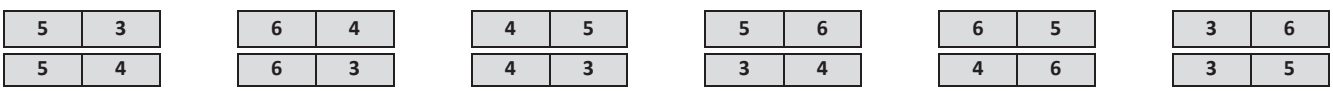

Fig. 3. Stimulus presentation in the PC-1 and PC-1×2 study
Rys. 3. Prezentacja materiału w eksperymencie PC-1 i PC-1×2

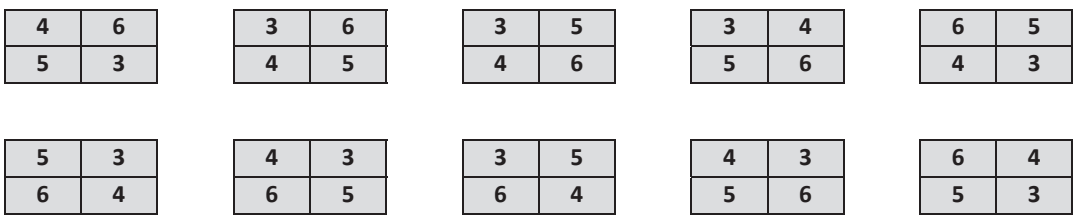

Fig. 4. Stimulus presentation in the PC-2×2 experiment
Rys. 4. Prezentacja materiału w eksperymencie w eksperymencie PC-2×2

Table 1. Percentage of correct votes in PC-1 experiment
Tabela 1. Procentowy udział poprawnych odpowiedzi w eksperymencie PC-1

Mbps	3	4	5	6
3	x	69.23	97.44	94.87
4	87.17	x	92.31	100
5	94.87	92.31	x	82.05
6	97.44	94.87	74.36	x

Table 2. Percentage of correct votes in PC experiment PC-1×2
Tabela 2. Procentowy udział poprawnych odpowiedzi w eksperymencie PC-1×2

Mbps	3	4	5	6
3	x	81.25	93.75	100
4	93.75	x	100	100
5	100	100	x	81.25
6	100	100	87.5	x

In all other cases, the level of correctness of observers is higher for PC-1×2. The number of pairs for which 100% of votes are correct has increased from one in the case of PC-1 to seven for PC-1×2.

The observers performed worst when the bitrate values were close to each other and close to the boundary values (highest or lowest) at the same time, namely the 3 & 4, 4 & 3, 5 & 6, and 6 & 5 Mbps pairs. The phenomenon was not observed for the 4 & 5 and 5 & 4 Mbps pairs. It is therefore likely that compression artefacts are still visible in the 4 Mbps video but disappear in the 5 Mbps one. In contrast, visual differences in the cases of the 3 & 4, 4 & 3, 5 & 6, and 6 & 5 Mbps pairs were already harder to notice. In most cases, the PC-1×2 test, where an observer viewed both images at the same time, gave a higher number of correct votes. This means that according to previous assumptions, the experiment was easier for participants and gave more accurate results.

The results from the PC-2×2 test are displayed in Fig. 6, which shows a general summary, for the whole study, of the total number of votes cast for specific bit streams considered to be the best (green) or the worst (red) quarters of the image.

As the observer was watching all four bitrate representations at the same time, all votes cast for 6 Mbps as the best part and all votes cast for 3 Mbps as the worst may be indexed as correct. Therefore the percentage of votes cast correctly is very high: 80.51% in the case of the best quality and 87.95% in case of the

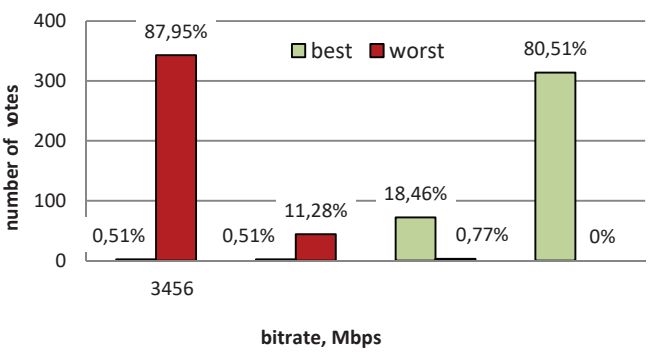

Fig. 6. Percentage of votes given to a quarter with the best (green) and the worst (red) quality with respect to the bitrate
Rys. 6. Procent głosów oddanych na ćwiartkę obrazu najlepszej (kolor zielony) i najgorszej (kolor czerwony) jakości w zależności od wielkości strumienia bitowego – większa wartość strumienia bitowego to lepsza jakość

worst. Next, in order of quality, 5 Mbps was denoted as the best by 18.46% of votes, while 4 Mbps was denoted as the worst by 11.28% of votes. A the same time there were two votes (0.51%) by observers indicating that the 4 Mbps and 3 Mbps streams were the best and three votes(0.77%) indicating that 5 Mbps streams were the worst. No observer voted for 6 Mbps as the worst. In general, it was easier for observers to identify quarters with the worst quality than those with the best. This shows, among other things, that humans focus primarily on visual distortions caused by excessive compression of the material. The better the image quality is, the more difficult it becomes to assess.

The other research question was whether the correctness of the vote depends on the quarter in which the material is displayed. Figure 7 shows the number of correct indications for each quarter of the screen, taking into account indications of both the best and the worst parts.

The correctness obtained in PC-2×2 was high for each quarter. The lowest observed correctness level was 75.21%, in this case for the best quarter in the lower right quarter. The highest observed accuracy level was 92.31% for the worst part in the upper right quarter. In all quarters of the screen there was a consistent percentage of correct indications with an average value of 80.77% for the best quarter and 88.04% for the worst (average calculated for three quarters of the screen, as there was no 3 Mbps sequence in the lower left corner). None of the image quadrants were privileged by observers. The number of correct votes cast for the worst part was slightly higher than the number of correct votes cast for the best. This also shows once again that it is easier for observers to judge a decrease in image quality than an improvement. It may be assumed that there is no problem focusing attention

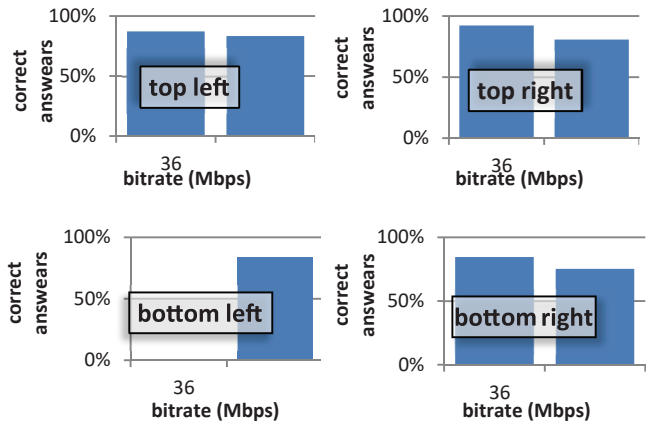

Fig. 7. Percentage of correct answers for the maximum (6 Mb/s) and the minimum (3 Mb/s) quality level depending on the frame quarter in which the stimulus was presented
Rys. 7. Procentowy udział poprawnych odpowiedzi dla maksymalnego (6 Mb/s) oraz minimalnego (3 Mb/s) poziomu jakości w zależności od ćwiartki kadru, w którym prezentowany był bodziec

on the image divided into four parts and voting for two of them after viewing the test sequence

4. Conclusion

Our idea of test material presentation made it possible to maintain more natural viewing conditions and to use a cheaper and more flexible experimental station. In the PC-1×2 experiment, limiting the image quality information to half of the full screen did not disturb the image quality assessment, but the duration of the test was halved. Simultaneous presentation of the two stimuli resulted in increases in the number of scores cast correctly. What is more, the human audience was able to recognize and correctly evaluate four quarters coded with different bitrates displayed in parallel. Obtaining such promi-

sing results encourages further research involving experiments with natural videos instead of rendered ones and preferably of ultra-high resolution.

Acknowledgement

This work was fully supported by the statutory funds of Institute of Metrology and Biomedical Engineering, Warsaw University of Technology.

Bibliography

1. Shi Y., Sun H., *Image and Video Compression for Multimedia Engineering*, Image Processing Series, (CRC Press, Boca Raton, 2008), DOI: 10.1201/9781420007268.
2. Moldovan A.-N., Ghergulescu I., Muntean C. H., *VQAMap: A Novel Mechanism for Mapping Objective Video Quality Metrics to Subjective MOS Scale,* "IEEE Transactions on Broadcasting", Vol. 62, No. 3, 2016, 610–627, DOI: 10.1109/tbc.2016.2570002.
3. ITU-T Recommendation P.910: *Subjective video quality assessment methods for multimedia applications*, 2008.
4. Kozamernik F., Steinmann V., Sunna P., Wyckens E., *SAMVIQ—A New EBU Methodology for Video Quality Evaluations in Multimedia*, "SMPTE Motion Imaging Journal", Vol. 114, No. 4, 2005, 152–160, DOI: 10.5594/j11535.
5. ITU-R Recommendation BT.500-13: 'Methodology for the subjective assessment of the quality of television pictures', 2012
6. Hoffmann H., Itagaki T., Wood D., Hinz T., Wiegand T., *A Novel Method for Subjective Picture Quality Assessment and Further Studies of HDTV Formats*, "IEEE Transactions on Broadcasting", Vol. 54, No. 1, 2008, 1–13, DOI: 10.1109/tbc.2008.916833.
7. Pinson M.H., Wolf S., *Comparing subjective video quality testing methodologies*, Visual Communications and Image Processing, 2003, DOI: 10.1117/12.509908.
8. Zeng K., Zhao T., Rehman A., Wang Z., *Characterizing perceptual artifacts in compressed video streams*, Proceedings of SPIE 9014, *Human Vision and Electronic Imaging XIX*, 90140Q (25 February 2014); DOI: 10.1117/12.2043128.

Percepcyjna metoda oceny jakości materiału wideo z użyciem dowolnej liczby obrazów odniesienia

Streszczenie: Artykuł prezentuje nową metodę oceny jakości materiału wideo. Koncepcja tej metody zakłada prezentację poddawanego ocenie materiału testowego równolegle z materiałem odniesienia, co ułatwia sformułowanie właściwej oceny. Nowatorskim pomysłem jest podział obrazu wideo w pełnej rozdzielczości na n-obrazową macierz. Każda komórka takiej macierzy może być zakodowana z innymi parametrami kompresji. W artykule zaprezentowano wyniki eksperymentów z zastosowaniem podziału na cztery części. Przeprowadzone analizy pokazują, że możliwe jest uzyskanie dokładniejszych wyników w krótszym czasie w porównaniu do klasycznej metody oceny.

Słowa kluczowe: kompresja informacji wizualnej, ocena jakości materiału wideo

Rafał Kłoda, PhD
rkloda@piap.pl

Graduated in the field of Automation and Robotics at the Faculty of Mechatronics, Warsaw University of Technology. In 2014 he became a doctor of science in automation and robotics from the same University. During the doctoral studies he worked on many aspects of video quality evaluation and algorithms that will be able to forecast mean opinion score from observers and use this knowledge to detect defects in compressed video in production process. He has wide research and teaching experience in video quali-ty evaluation and video processing. Experienced in statistical analyses, development and validation of IT systems also for security applications. Involved in development of data fusion and visualization for military applications. In recent years he has been in-volved in the development of novel measurement methods (especially to video quality evaluation), and novel IT systems for e-learning applications. His current research interests include expert systems, modeling and visualization.

Anna Ostaszewska-Liżewska, PhD
a.ostaszewska@mchtr.pw.edu.pl

Graduated (2004) from the Warsaw University of Technology (WUT). She is an assistant professor at the Institute of Metrology and Biomedical Engineering of WUT. Her current research interests include: virtual reality and computational intelligence methods in compressed video quality evaluation.