Perceived Quality of Full HD Video - Subjective Quality Assessment

Juraj BIENIK, Miroslav UHRINA, Martin VACULIK, Tomas MIZDOS

Department of Telecommunications and Multimedia, Faculty of Electrical Engineering, University of Zilina,
Univerzitna 8215/1, 010 07 Zilina, Slovakia

juraj.bienik@fel.uniza.sk, miroslav.uhrina@fel.uniza.sk, martin.vaculik@fel.uniza.sk, t.mizdos@gmail.com

DOI: 10.15598/aeee.v14i4.1831

Abstract. *In recent years, an interest in multimedia services has become a global trend and this trend is still rising. The video quality is a very significant part from the bundle of multimedia services, which leads to a requirement for quality assessment in the video domain. Video quality of a streamed video across IP networks is generally influenced by two factors – transmission link imperfection and efficiency of compression standards. This paper deals with subjective video quality assessment and the impact of the compression standards H.264, H.265 and VP9 on perceived video quality of these compression standards. The evaluation is done for four full HD sequences, the difference of scenes is in the content – distinction is based on Spatial (SI) and Temporal (TI) Index of test sequences. Finally, experimental results follow up to 30 % bitrate reducing of H.265 and VP9 compared with the reference H.264.*

Keywords

ACR, DSCQS, H.264/AVC, H.265/HEVC, subjective assessment, video quality, VP9.

1. Introduction

In the last years the demand for multimedia services is still rising and the amount of video streaming has grown more and more especially. Due to the quantity of video streams and the requirement of bandwidth the need for developing effective compression has occurred.

The paper is divided as follows. The first part of the article provides rudimentary information about the mentioned compression standards. In the second part, subjective quality metrics and the process of quality assessment used in experimental measurements are described. The last part of this article deals with the experiment results and conclusions which stem from the measurements of perceived video quality by the observers.

Nowadays many compression standards are being introduced, e.g. H.265/HEVC, VP9, DAALA and the video quality of them was tested [1], [2], [3] and [4]. Each of these mentioned standards indicate a high level of compression. Their comparison in terms of subjective quality is the aim of this paper. Quality comparison of compression standard is very important to providers of video services and end users as well.

2. Compression Standards

The Advanced Video coding known as H.264/AVC (MPEG-4 Part 10) is the oldest of the mentioned compression standards (approved in 2003), but globally still most used. The versatility of this standard provides a wide range of applications from video in smartphones to TV broadcasting and multimedia content on Blu-ray discs.

The High Efficiency Video Coding known as H.265/HEVC (approved in January 2013) is the most recent joint video cooperation result of the ITU-T Video Coding Experts Group (VCEG) and the ISO/IEC Moving Picture Experts Group (MPEG) standardization organizations. Collaboration of these groups and participation on this project is known as the Joint Collaborative Team on Video Coding (JCT-VC). H.265 is the successor to the very popular H.264 standard. The basic features and structure of H.265 stay the same as its predecessor, but it is considered to contain many significant improvements which make video compression more effective [5], [6] and [7].

VP9 is the WebM Project's next-generation open video codec and VP9 is the direct successor of VP8, which was the biggest competitor to H.264. WebM is an open, royalty-free media file format. VP9 was approved in June 2013. The most prominent features of WebM can be considered the openness, innovation and optimisation for the web. The main aim of the WebM Project is to speed up the pace of video compression innovation (i.e. to get better and faster). VP9 was enabled by default in the Google Chrome Dev channel [8].

3. Video Quality Assessment

Video quality assessment can be divided into two types of methods. The first one is objective quality and the second is subjective quality. Objective quality methods consist of computation methods called "metrics" which are based on signal analyses of pictures. Metrics produce values that represent quality. The most used objective metrics are Peak Signal-to-Noise Ratio (PSNR), Video Quality Metric (VQM) and Structural Similarity Index (SSIM). The oldest objective metric is PSNR [9], but it's still very popular and often used because it can be computed very easily and quickly. The SSIM metric measures three components (the similarity of luminance, contrast and structure) and combine them to a value in the range from 0 to 1, where 0 is the worst and 1 is the best quality [10]. The VQM metric computes the visibility of artefacts expressed in the DCT domain. The output value represents the amount of distortion with the best quality indicated by a value close to zero [11].

Subjective quality assessment is based on the vote of human (observers), quantify perceived video quality using discrete values from a certain range (scale depending on chosen method). The biggest benefit of subjective quality assessment is the credibility of the results - objective methods do not achieve such accuracy of results (they are based only on a model of perceived quality) and values from metrics are only approximations of real video quality. The drawback of subjective methods is that it is time-consuming and human resources are needed.

Most used subjective methods are:

- DSIS - Double Stimulus Impairment Scale also known as Degradation Category Rating (DCR).
- DSCQS - Double Stimulus Continuous Quality Scale.
- SSCQE - Single Stimulus Continuous Quality Evaluation.
- ACR - Absolute Category Rating also known as Single Stimulus (SS).
- SDSCE - Simultaneous Double Stimulus for Continuous Evaluation [9] and [10].

Procedures and conditions for subjective quality assessment are defined in ITU-R BT.500-13 [12]. This recommendation defines that a minimum of 15 observers should be used to achieve reliable results. They should be non-experts for the assessment of video quality and their normal work is not experienced assessors. The count of assessors depends on the sensitivity and reliability of the test procedure. Before the start of a testing session assessors should be familiar with many factors, for example the methods of assessment, grading scale, the type of impairments, the timing (duration of training, test and reference sequences, time for voting) and so on [10].

The whole session should not take longer than 30 minutes. Before the first test session there should be 3 to 5 sequences shown to stabilize the opinion of the observer. The order used for the presentation should be random, but the test condition order should be set so that any effects on the grading of fatigue or adaptation are balanced out in all sessions uniformly. To check the coherence there should be some presentations repeated from session to session [9].

After the test session the calculation of Mean Opinion Score (MOS) is done:

$$\bar{u}_{jkr} = \frac{1}{N}\sum_{i-1}^{N} u_{ijkr}, \tag{1}$$

where u_{ijkr} is the score of assessor i for test condition j, sequence k, repetition r and N stands for a number of accessors.

Finally, the 95 % confidence interval, which is derived from standard deviation and size of each sample is computed. It is given by:

$$\delta_{jkr} = 1.96 \cdot \frac{S_{jkr}}{\sqrt{N}}, \tag{2}$$

where:

$$S_{jkr} = \sqrt{\sum_{i=1}^{N} \frac{(u_{jkr} - u_{ijkr})^2}{(N-1)}} \quad [13]. \tag{3}$$

In our experiments, DSIS and ACR methods were used.

3.1. The Double-Stimulus Impairment Scale Method - DSIS

This method consists of pair a pair sequences. The first sequence is unimpaired (the reference) and the second sequence is impaired due to compression (the test). Order of sequences is still the same (Fig. 1) and the assessor is acquainted with this order.

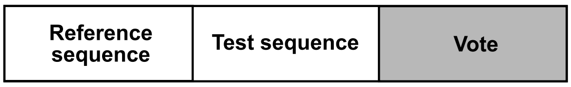


Fig. 1: DSIS - order of sequences.

Assessors see the reference sequence first, keeping that in mind, the test sequence follows and then the assessor rates the grade of impairment (difference) between the reference and test sequence with a value from five-grade scale, where:

- 5 = imperceptible,
- 4 = perceptible, but not annoying,
- 3 = slightly annoying,
- 2 = annoying,
- 1 = very annoying [9], [10] and [13].

3.2. The Absolute Category Rating Method - ACR

Unlike the previous method, ACR (also known as Single Stimulus method - SS) consists only of degraded sequences, without a reference sequence (Fig. 2). The assessor is evaluating the level of quality with the value from the five-level grading scale, where:

- 5 = excellent,
- 4 = good,
- 3 = fair,
- 2 = poor,
- 1 = bad [9], [10] and [13].

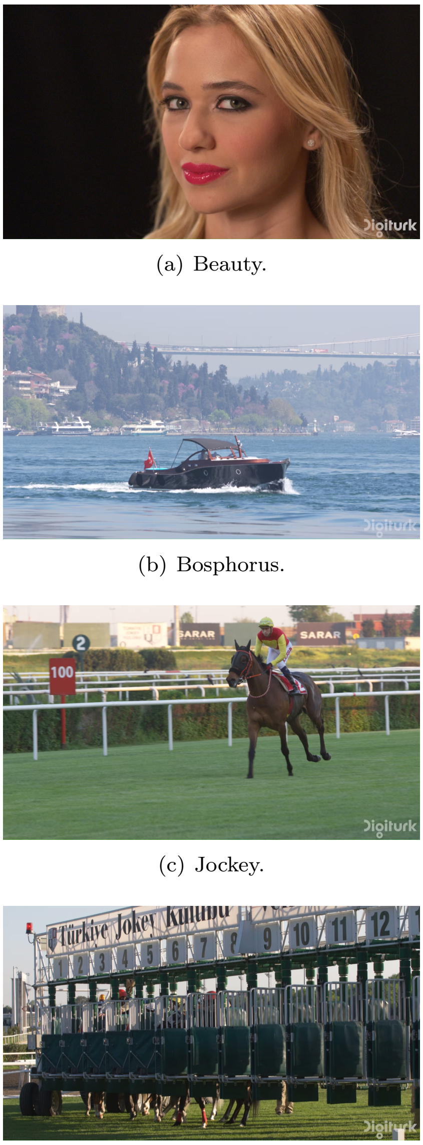
Test sequence
Vote

Fig. 2: Traffic model of multiple nodes for simulations.

4. Measurements and Experimental Results

In our experiments four types of test sequences with different content were used:

- "Beauty" (Fig. 3(a)) – a detail of a female's face, her hair is slowly blowing in the wind on the static black background.
- "Bosphorus" (Fig. 3(b)) – a boat sailing in the Bosphorus strait with a huge bridge in the background, the camera panning from left to right - one object with slow motion.
- "Jockey" (Fig. 3(c)) – a running horse with a rider, the camera panning from left to right – one object with quick motion.
- "ReadySteadyGo" (Fig. 3(d)) – a horserace, horses with jockeys are competing, camera panning from left to right, several objects with quick motion.



(a) Beauty.

(b) Bosphorus.

(c) Jockey.

(d) ReadySteadyGo.

Fig. 3: Test sequences.

All sequences were in full HD resolution (1920 × 1080 pixels), the aspect ratio 16:9 and framerate of 60 fps (frames per second). The length of each sequence was 10 seconds.

Since the compression difficulty is directly related to the spatial and temporal information of a sequence, regarding [13] the Spatial Information (SI) and Temporal Information (TI) of all sequences using the Mitsu tool [14] were calculated. The results are shown in the Tab. 3. According to results the spatial-temporal information plane was drawn (Fig. 4).

The measurement process consists of the following steps:

- First, all sequences in uncompressed format (*.yuv) from [15] were downloaded.
- Afterwards, they were encoded to the H.264, H.265 and VP9 compression standards using FFmpeg [16]. Target bitrates of all sequences were 1, 2, 3, 5, 7, 10 and 15 Mbps, size of GOP = 30, number of B-frame = 5.
- Then, compressed sequences in container *.mp4 and *.mkv were encoded to raw format with *.avi container.
- From all sequences playlists were compiled – the order of sequences with target bitrates for DSIS and ACR can be seen in Tab. 1 and Tab. 2.
- All compression standards were assessed by a group of observers consisting of 30 people (in total 3 groups with 90 observers). Observers watched a playlist on a 42" television, all test sequences in 7 target bitrates and assessed with DSIS method first and afterwards with ACR method (observers didn't know the sequences bitrate order). Finally, the observers assessed the video sequences with a value from the grading scale.

Tab. 1: Sequence order with target bitrate for DSIS method.

Sequence number / Scene (Mbps)	1	2	3	4	5	6	7
Beauty	10	1	15	2	7	3	5
Bosphorus	1	7	10	2	5	15	3
Jockey	3	1	7	15	5	10	2
ReadySreadyGo	10	5	3	1	15	7	2

Tab. 2: Sequence order with target bitrate for ACR method.

Sequence number / Scene (Mbps)	1	2	3	4	5	6	7
Beauty	5	15	7	10	2	3	1
Bosphorus	2	73	1	5	15	7	10
Jockey	5	2	15	3	7	1	10
ReadySreadyGo	10	5	2	1	7	15	3

Detail information about observers from evaluation groups 1, 2 and 3 are specified in Tab. 4.

Tab. 3: Spatial Index (SI) and Temporal Index (TI) of test sequences.

	Beauty	Bosphorus	Jockey	ReadySteadyGo
SI	24.33	36.51	28.11	68.81
TI	7.15	4.68	20.47	28.58

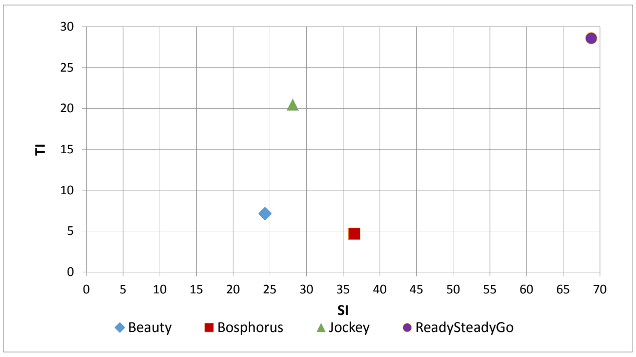


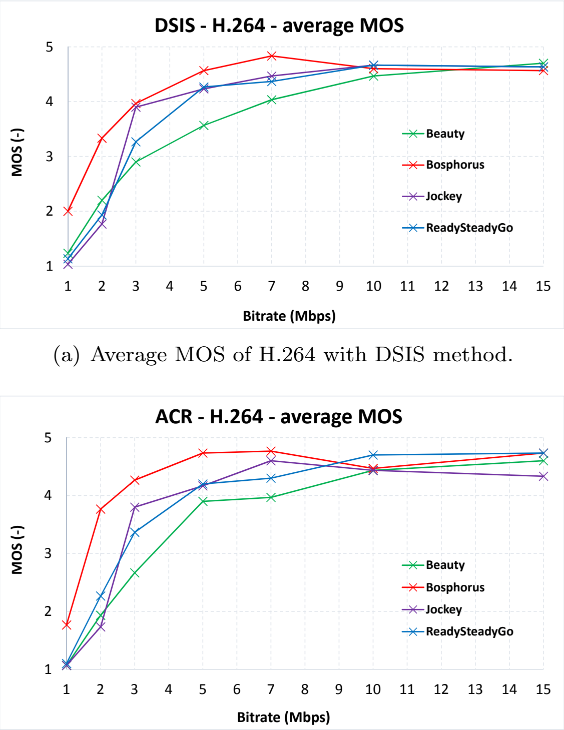
Fig. 4: Spatial-temporal plot of all test sequences.

Tab. 4: Information about observers individual group.

	Count of assessors (-)	Count of men (-)	Count of women (-)	Average age of assessors (years)	Average age of men (years)	Average age of women (years)
Group1 H.264 assessment	30	22	8	24.067	24.273	23.5
Group2 H.265 assessment	30	28	2	26.233	26.214	26.5
Group3 VP9 assessment	30	23	7	28.7	28.609	29

From the assessment tables we computed averages of MOS values for each compression standard in target bitrates 1, 2, 3, 5, 7, 10 and 15 Mbps. According to the results from subjective assessment graphs, which showed average MOS value of compression standards for DSIS and ACR methods, (Fig. 5(a), Fig. 5(b), Fig. 6(a), Fig. 6(b), Fig. 7(a) and Fig. 7(b)) were created. In graphs for all measured values 95 % confidence interval were depicted to determine quality saturation (quality threshold). To find out quality saturation value, overlay of lines from 95 % confidence interval were used. This value corresponds to a tradeoff between perceived quality and bitrate; that it is not necessary to increase the bitrate, influence for grow of quality is minimal. Coding efficiency comparison for used compression standards in the same scene and with used DSIS and ACR methods is shown in the graphs (Fig. 8(a), Fig. 8(b), Fig. 9(a), Fig. 9(b), Fig. 10(a), Fig. 10(b), Fig. 11(a) and Fig. 11(b)), where there is an important portion of MOS curves in range of bitrate from 1 to 7 Mbps (till quality threshold).

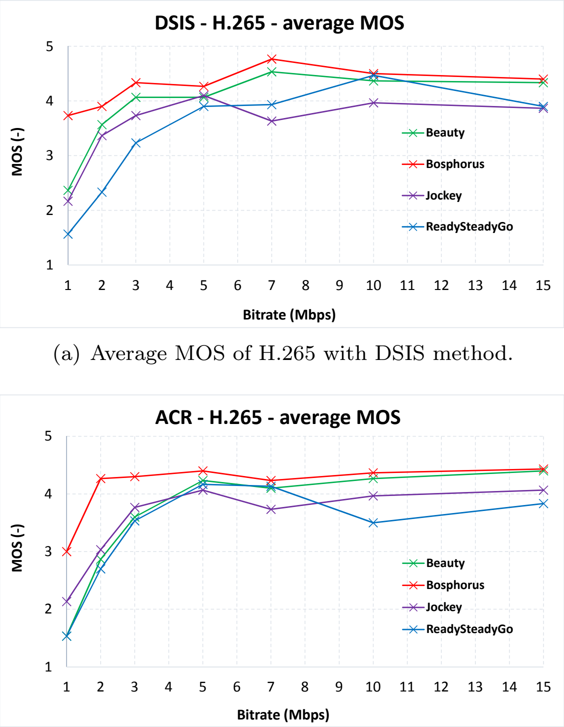
Generally, the observers evaluated the scene "Bosphorus" as a best one, because in this scene there is not much motion and it contains a big amount of structural changes, which are harder perceived by observers. Vice versa, the worst quality indicates quick motion scenes in both methods.



(a) Average MOS of H.264 with DSIS method.

(b) Average MOS of H.264 with ACR method.

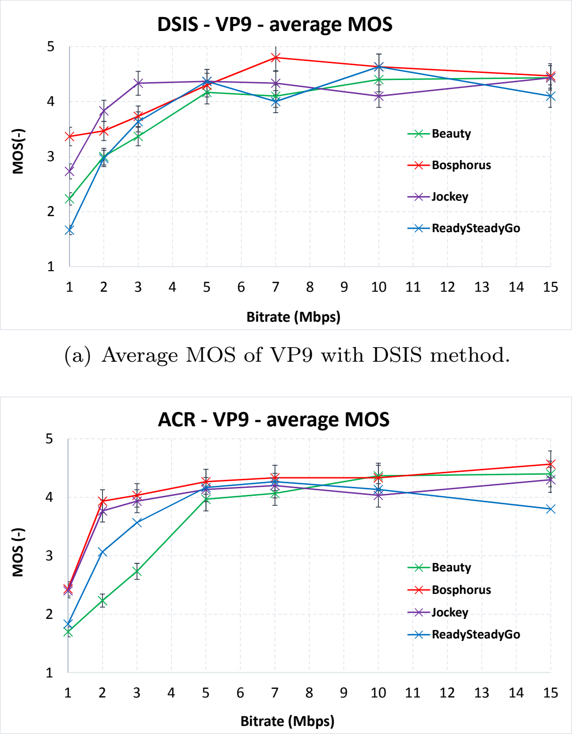
Fig. 5: Average MOS of H.264 compression standard.



(a) Average MOS of H.265 with DSIS method.

(b) Average MOS of H.265 with ACR method.

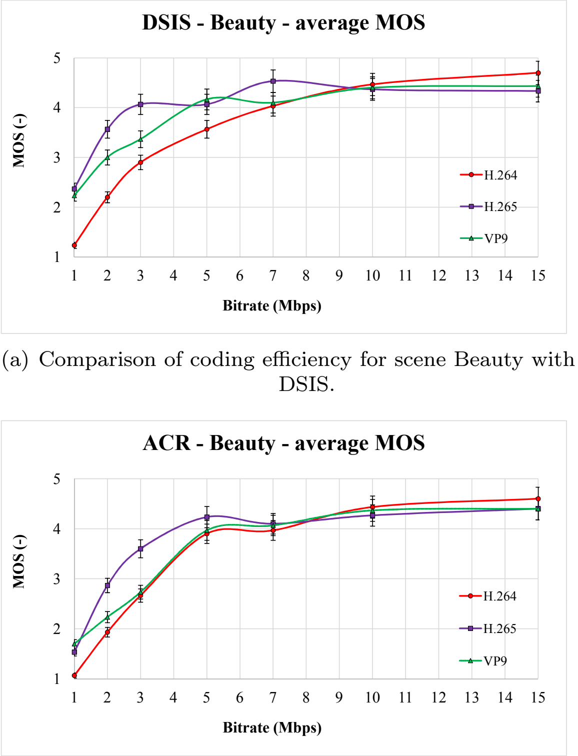
Fig. 6: Average MOS of H.265 compression standard.



(a) Average MOS of VP9 with DSIS method.

(b) Average MOS of VP9 with ACR method.

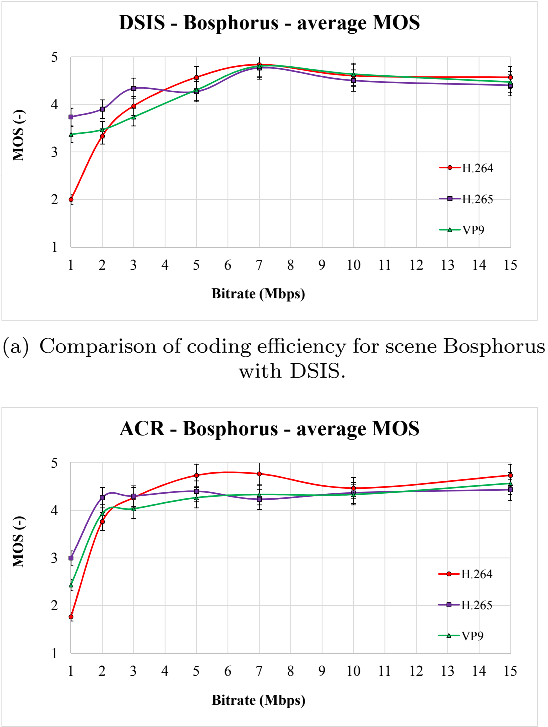
Fig. 7: Average MOS of VP9 compression standard.



(a) Comparison of coding efficiency for scene Beauty with DSIS.

(b) Comparison of coding efficiency for scene Beauty with ACR.

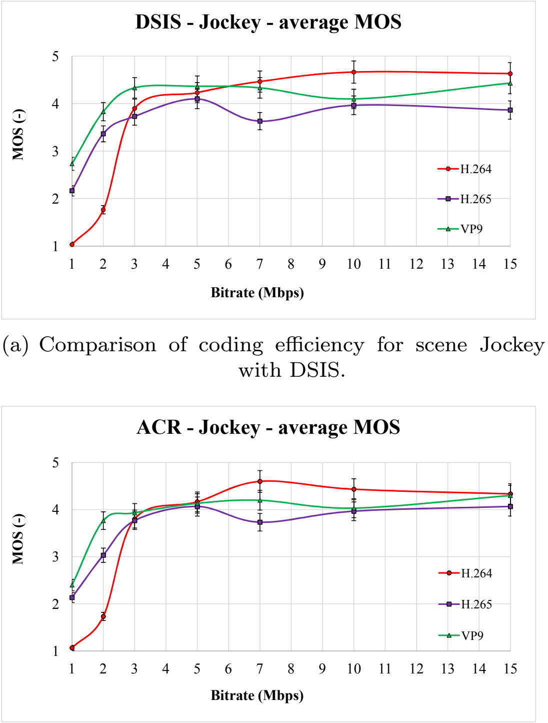
Fig. 8: Comparison of coding efficiency for scene Beauty with used DSIS and ACR method.



(a) Comparison of coding efficiency for scene Bosphorus with DSIS.

(b) Comparison of coding efficiency for scene Bosphorus with ACR.

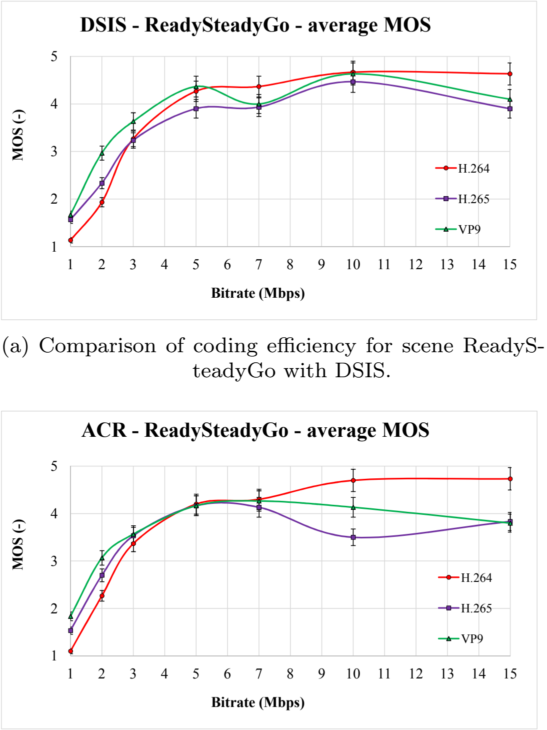
Fig. 9: Comparison of coding efficiency for scene Bosphorus with used DSIS and ACR method.



(a) Comparison of coding efficiency for scene Jockey with DSIS.

(b) Comparison of coding efficiency for scene Jockey with ACR.

Fig. 10: Comparison of coding efficiency for scene Jockey with used DSIS and ACR method.



(a) Comparison of coding efficiency for scene ReadySteadyGo with DSIS.

(b) Comparison of coding efficiency for scene ReadySteadyGo with ACR.

Fig. 11: Comparison of coding efficiency for scene ReadySteadyGo with used DSIS and ACR method.

5. Conclusion

This paper dealt with evaluating the impact of the H.264/AVC, H.265/HEVC and VP9 compression standards on the perceived video quality using selected subjective metrics. The target of this paper was to research how non-expert observers perceived and evaluated the video quality affected by the bitrate. The evaluation was done for four types of Full HD sequences with different content. From the graphs we should state that the threshold of the perceived quality of the H.265 a VP9 compression standards is close to 5 Mbps bitrate and quality saturation of H.264 in approximately 7 Mbps. This fact leads to the conclusion that there is no need for providers to use higher bitrates in streaming than this threshold, so they can save space in the transmission chain and use it for other channels or services. It follows that both new compression standards (VP9 and H.265) outperformed H.264 and exhibit a higher level of compression, mainly in lower bitrates till 7 Mbps. Over quality threshold exhibits H.264 higher performance than newer compression standards. The reason of this fact should be that H.264 was developed exactly for full HD resolution, vice versa H.265 and VP9 were designed mainly for 4K resolution video. In the near future we plan to extend the analysis of the impact of H.265/HEVC and VP9 compression standards with Ultra HD resolution on video quality using subjective metrics.

References

[1] GROIS, D., D. MARPE, A. MULAYOFF, B. ITZHAKY and O. HADAR. Performance Comparison of H.265/MPEG - HEVC, VP9, and H.264/MPEG-AVC Encoders. In: *Picture Coding Symposium (PCS)*. San Jose: IEEE, 2013, pp. 394–397. ISBN 978-1-4799-0292-7. DOI: 10.1109/PCS.2013.6737766.

[2] RAO, K. R. Video coding standards: AVS China, H.264/MPEG-4 part 10, HEVC, VP9, DIRAC and VC-1. In: *Signal Processing: Algorithms, Architectures, Arrangements, and Applications*. Poznan: IEEE, 2013, pp. 334–340. ISBN 978-83-62065-17-2.

[3] KIM, I. K., S. LEE, Y. PIAO and J. CHEN. Coding efficiency comparison of new video coding standards: HEVC vs VP9 vs AVS2 video. In: *IEEE International Conference on Multimedia and Expo Workshops*. Chengdu: IEEE, 2014, pp. 1–6. ISBN 978-1-4799-4717-1. DOI: 10.1109/ICMEW.2014.6890700.

[4] RERABEK, M., P. HANHART, P. KORSHUNOV and T. EBRAHIMI. Quality evaluation of HEVC and VP9 video compression in real-time applications. In: *7th International Workshop on Quality of Multimedia Experience (QoMEX)*. Costa Navario: IEEE, 2015, pp. 1–6. ISBN 978-1-4799-8958-4. DOI: 10.1109/QoMEX.2015.7148088.

[5] SULLIVAN, G. J., J. R. OHM, W. J. HAN and T. WIEGAND. Overview of the High Efficiency Video Coding (HEVC) Standard. *IEEE Transactions on Circuits and Systems for Video Technology*. 2012, vol. 22, no. 12, pp. 1649–1668. ISSN 1051-8215. DOI: 10.1109/TCSVT.2012.2221191.

[6] ITU/T H.265 High efficiency video coding. *ITU-T Recommendation* [online]. 2015. Available at: http://www.itu.int/ITU-T/recommendations/rec.aspx?rec=12455&lang=en.

[7] ITU/T H.264 Advanced video coding for generic audiovisual services. *ITU-T Recommendation* [online]. 2016. Available at: https://www.itu.int/rec/T-REC-H.264-201602-S/en.

[8] VP9 Compression standards. *WebM Project*. [online]. 2013. Available at: https://www.webmproject.org/vp9/.

[9] WINKLER S. *Digital Video Quality: Vision Models and Metrics*. Hoboken: John Wiley and Sons, 2005. ISBN 0-470-02404-6.

[10] WU, H. R. and K. R. RAO. *Digital Video Image Quality and Perceptual Coding*. Boca Raton: CRC Press, 2006. ISBN 0-8247-2777-0.

[11] LOKE, M. W., E. P. ONG, W. LIN, Z. LU and S. YAO. Comparison of Video Quality Metrics on Multimedia Videos. In: *International Conference on Image Processing*. Atlanta: IEEE, 2006, pp. 457–460. ISBN 1-4244-0480-0. DOI: 10.1109/ICIP.2006.312492.

[12] BT.500-13 - Methodology for the subjective assessment of the quality of television pictures. *ITU-R Recommendation* [online]. 2013. Available at: https://www.itu.int/rec/R-REC-BT.500.

[13] P.910, Subjective video quality assessment methods for multimedia applications. *ITU-T Recommendation* [online]. 2008. Available at: https://www.itu.int/rec/T-REC-P.910-200804-I/en.

[14] ROMANIAK, P., L. JANOWSKI, M. LESZCZUK and Z. PAPIR. Perceptual quality assessment for H.264/AVC compression. In: *IEEE Consumer Communications and Networking Conference (CCNC)*. Las Vegas: IEEE, 2006, pp. 597–602. ISBN 978-1-4577-2071-0. DOI: 10.1109/CCNC.2012.6181021.

[15] Test Sequences. *Ultra Video Group*. [online]. 2016. Available at: http://ultravideo.cs.tut.fi/#testsequences.

[16] FFmpeg tool. *The FFmpeg project*. [online]. 2016. Available at: https://www.ffmpeg.org/.

About Authors

Juraj BIENIK was born in 1987 in Zilina, Slovakia. He received his M.Sc. in Telecommunications at the Department of Telecommunications and Multimedia, Faculty of Electrical Engineering, at the University of Zilina in 2012. Nowadays he is a Ph.D. student at the same department. His research interests include audio and video signal processing, functionality and optimisation of networks and video quality assessment.

Miroslav UHRINA was born in 1984 in Zilina, Slovakia. He received his M.Sc. and Ph.D. degrees in Telecommunications at the Department of Telecommunications and Multimedia, Faculty of Electrical

Engineering, at the University of Zilina in 2008 and 2012, respectively. Nowadays he is an assistant professor at the same department. His research interests include audio and video compression, video quality assessment, TV broadcasting and IP networks.

Martin VACULIK was born in 1951. He received his M.Sc. and Ph.D. in Telecommunications at the University of Zilina, Slovakia in 1976 and 1987 respectively. In 2001 he was habilitated as associate professor of the Faculty of Electrical Engineering at the University of Zilina in the field of Telecommunications. Currently he works as a head of Department of Telecommunications and Multimedia. His interests cover switching and access networks, communication network architecture, audio and video applications.

Tomas MIZDOS was born in 1993 in Poprad, Slovakia. He received his B.Sc. degrees in Multimedia technologies at the Department of Telecommunications and Multimedia, Faculty of Electrical Engineering, at the University of Zilina in 2015. Nowadays he is M.Sc. student at the same department. His main area of interest is functionality and quality of multimedia services.