Data in Brief 54 (2024) 110458

Contents lists available at ScienceDirect

Data in Brief

journal homepage: www.elsevier.com/locate/dib

Data Article

Video dataset containing video quality assessment scores obtained from standardized objective and subjective testing

Jaroslav Frnda[a,*], Marek Durica[a], Jerry Chun-Wei Lin[b], Philippe Fournier-Viger[c]

[a] Department of Quantitative Methods and Economic Informatics, Faculty of Operation and Economics of Transport and Communication, University of Zilina, 01026 Zilina, Slovakia
[b] Faculty of Automatic Control, Electronics and Computer Science, Department of Distributed Systems and IT Devices, Silesian University of Technology, Poland
[c] College of Computer Science and Software Engineering, Shenzhen University, Shenzhen 518060, China

ARTICLE INFO

Article history:
Received 27 February 2024
Revised 2 April 2024
Accepted 16 April 2024
Available online 23 April 2024

Dataset link: Video dataset for coding efficiency and image quality analysis containing results from standardized objective and subjective assessment methods. (Original data)

Keywords:
H.265
H.264
Packet loss
Mean opinion score
SSIM
VMAF

ABSTRACT

This paper presents a dataset comprising 700 video sequences encoded in the two most popular video formats (codecs) of today, H.264 and H.265 (HEVC). Six reference sequences were encoded under different quality profiles, including several bitrates and resolutions, and were affected by various packet loss rates. Subsequently, the image quality of encoded video sequences was assessed by subjective, as well as objective, evaluation. Therefore, the enclosed spreadsheet contains results of both assessment approaches in a form of MOS (Mean Opinion Score) delivered by the absolute category ranking (ACR) procedure, SSIM (Structural Similarity Index Measure) and VMAF (Video Multimethod Assessment Fusion). All assessments are available for each test sequence. This allows a comprehensive evaluation of coding efficiency under different test scenarios without the necessity of real observers or a secure laboratory environment, as recommended by the ITU (International Telecommunication Union). As there is currently no standardized mapping function between the results of subjective and objective methods, this dataset can also be used to design and verify

* Corresponding author.
E-mail address: jaroslav.frnda@uniza.sk (J. Frnda).

https://doi.org/10.1016/j.dib.2024.110458

experimental machine learning algorithms that contribute to solving the relevant research issues.

Specifications Table

Subject	Multimedia, Computer Networks and Communications
Specific subject area	Digital video broadcasting (DVB), Internet Protocol Television (IPTV), QoS and QoE analysis of video streaming
Type of data	Short 10 s length video sequences in .ts format Table (.xlsx format)
Data collection	The uncompressed UHD (Ultra High Definition) video sequences in raw YUV format were encoded using the FFmpeg tool in both H.264 and H.265 video compression standards with different parameters affecting the visual quality. The streaming testbed consisted of FFmpeg as the streaming server and the VLC player software as client for capturing the video stream. Various test scenarios with different packet loss rates were created using the Clumsy 0.3 software.
Data source location	University of Zilina, city of Zilina, Slovakia Owner of raw uncompressed original YUV video sequences is SJTU Media Lab, Shanghai Jiao Tong University, Shanghai, China
Data accessibility	Repository name: Mendeley Data Data identification number: 10.17632/35735kfjnm.1 Direct URL to data: https://data.mendeley.com/datasets/35735kfjnm/1 Raw original YUV video sequences are available here: https://medialab.sjtu.edu.cn/post/sjtu-4k-video-sequences/

1. Value of the Data

- Subjective video tests best reflect the end-user's (customer's) opinion of the picture quality provided. This allows the service operator (e.g. IPTV provider) to determine whether the quality is adequate or if customers complain about the delivered quality of service. However, subjective testing requires real observers, and it cannot be performed in real time, which makes this process time consuming. Therefore, if IPTV providers want to continuously monitor the service quality level they offer, which is typically affected by network behavior such as network congestion and packet loss, they must rely solely on objective methods of image quality evaluation (QoS concept) that can run in real-time. However, there is currently no standardized mapping function for interconnecting subjective and objective assessments. Our dataset could be used to design such a mapping function.
- For research groups working on a new video coding format, it is essential to understand how the proposed compression techniques and data network issues will affect the perceived image quality. Therefore, such datasets can be used to benchmark new codecs against existing video formats and assess the effectiveness of their designs. By analyzing how changes in encoding parameters affect quality scores, designers can refine their algorithms to achieve better compression while maintaining high visual quality.
- Hence, the main benefit of our dataset is that each testing video sequence is paired with both subjective and objective scores. Therefore, the dataset can be used to not only validate experimental mapping functions based on machine learning algorithms but also to design new full reference, as well as no-reference, objective methods for video quality assessment.
- The subjective testing followed all relevant ITU recommendations, and all individual ratings are presented in tables. The test scenarios represent the most common encoding settings used today.

2. Background

The primary goal of creating this dataset was to produce and evaluate video sequences of varying image quality. Short video sequences were encoded using different settings, such as codec, scene type, resolution, and bitrate, and they were affected by different packet loss rates. The obtained score can help to understand how different encoded profiles are robust to video quality deterioration caused by packet loss, and how to translate the results received from continuous monitoring of network parameters (QoS) to subjective perceived video quality (QoE). As the QoS concept is objective, it can be easily used to measure network parameters, such as the bitrate, the delay, the jitter and the packet rate of the IPTV provider's network. On the other hand, subjective testing (QoE) reflects the personal opinion of the end users (customers), thus the subjective evaluation is more important for IPTV companies than monitoring QoS parameters. Although the ITU has developed three models [1] that can predict subjective ratings based on reduced video data such as resolution, codec type, or bandwidth, these models do not account for degradations caused by packet loss, which can result in slicing, freezing, or blockiness in the picture. Predicting subjective quality rankings from objective metrics can help service providers allocate resources and optimize content delivery strategies based on user preferences. For instance, they can adjust adaptive bitrate streaming, which encodes a single video at multiple bitrates and resolutions, creating a so-called bitrate ladder. Each stream represents the same content but at different quality levels, allowing the delivery stream to be adapted to different network conditions and device capabilities. This is why finding the proper mapping function is still under research [2,3]. In addition, it is essential, not only for service providers but also for regulatory authorities (IPTV is subject to regulation), to know whether customer complaints are legitimate. Compared to recent datasets such as LIVE-NFLX-II [4] or the UVG dataset [5], our dataset contains test video sequences affected by packet loss. Additionally, we have not only measured objective metrics but also conducted subjective tests. Hence, this dataset offers a large number of test scenarios and subjective ratings that could help to find the relationship between subjective and objective scores.

3. Data Description

The dataset includes video sequences encoded from raw YUV format in the following settings (a combination of all settings allowed us to collect 700 testing video sequences):

- Resolution: HD, FullHD and UHD (4 K)
- Bitrate: 5, 10 and 15 Mbit/s
- Codec: H.264 (AVC) and H.265 (HEVC)
- Number of test scenes: 6
- Digital container format: .ts (MPEG transport stream)
- Packet loss rate: 0.1 %, 0.2 %, 0.3 %, 0.5 %, 0.75 % and 1 %

Six different types of scenes were selected based on their spatial-temporal information. The spatial (spatial detail in a picture) and temporal (amount of temporal change of the video sequence) information (SI/TI) used for characterization of the test sequences is calculated according to the guidelines provided in Recommendation ITU-T P.910 [6]. It is important to ensure that the set of test scenes covers the full range of SI and TI. Fig. 1 shows the SI-TI values for the following reference video sequences [7]:

- *Construction Field*—A construction vehicle is surrounded by buildings under construction. It contains dynamic objects such as an excavator and walking workers. The scene is captured as a static shot.
- *Runners* video shows a marathon race, also captured as a static shot, with dynamic movement of racers.

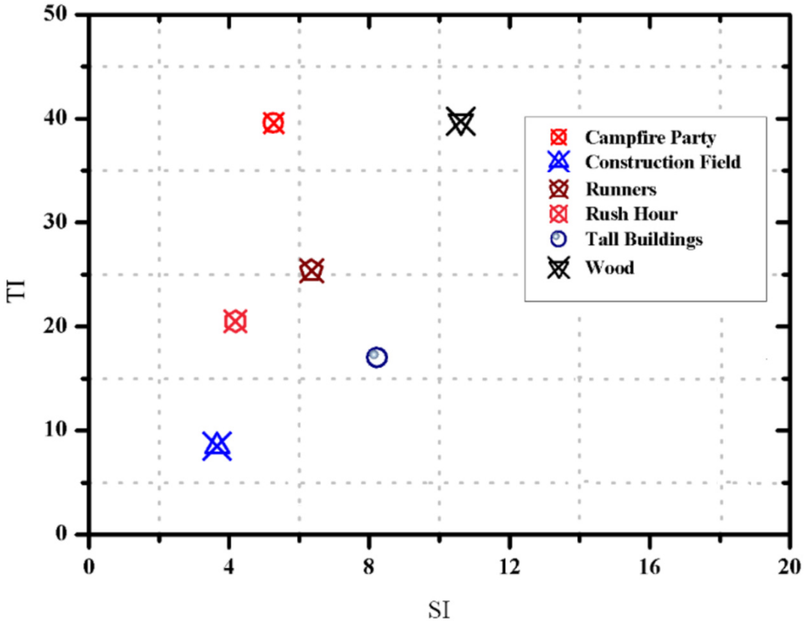


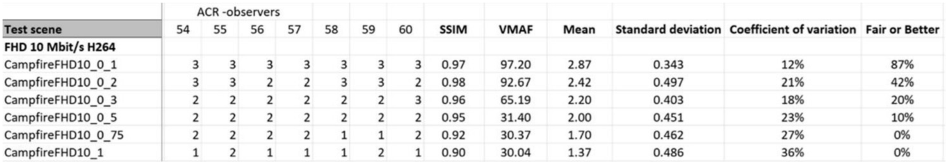
Fig. 1. Spatial information (SI) and temporal information (TI) values for UHD video sequences.

- *Wood* shows a woodland scene, with the camera moving dynamically from left to right, accelerating in sequence. Highest scores for spatial and temporal information.
- *Campfire Party* depicts a night scene with people gathered around a bonfire. The flame moves fast.
- *Tall Building* shows the tall buildings of Shanghai with a slow moving camera.
- *Rush Hour* shows students moving to canteen or dormitory after class.

The Absolute Category Rating (ACR) method was used to evaluate each tested video sequence, as described in the mentioned ITU recommendation [6]. ACR is also known as the single stimulus method, as observers do not compare the tested video sequence with the original one. Instead, they evaluate it based solely on their personal preferences. Each testing sequence should be approximately 10 s long, and breaks are required after each half hour of testing. The ACR method uses the Mean Opinion Score (MOS) which is represented by a five-point rating scale (1 - poor, 5 - excellent). This method reflects real-world situation where end-users (customers) do not have access to reference sequence for comparison. As a multimedia player, Media Player Classic - Home Cinema (MPC–HC) 64-bit version was used.

To have the best-known correlation methods with human perception, we chose two objective metrics. The first one is the SSIM (introduced in 2004 by prolific researchers Wang, Bovik et al.) [8], and the second one is VMAF (a collaboration between Netflix and several universities, published in 2016) [9]. Both SSIM and VMAF are full-reference metrics, meaning that they require a reference sequence to calculate a score. These metrics use mathematical and statistical methods to simulate the human visual system, taking into account factors such as higher sensitivity to contrast rather than to absolute luminance, different color sensitivity or blur intensity. SSIM uses a scale from 0 to 1, where 1 represents two identical sequences, while VMAF uses a scale from 0 to 100 (higher is better).

Individual scores are represented in MOS scale and are subject to statistical analysis. Statistical analysis consists of calculation of mean, standard deviation and coefficient of variation (CoV) [10]. CoV tells us if the mean can be used as a representative statistical parameter. If the CoV is



	ACR -observers												
Test scene	54	55	56	57	58	59	60	SSIM	VMAF	Mean	Standard deviation	Coefficient of variation	Fair or Better
FHD 10 Mbit/s H264													
CampfireFHD10_0_1	3	3	3	3	3	3	3	0.97	97.20	2.87	0.343	12%	87%
CampfireFHD10_0_2	3	3	2	2	3	3	2	0.98	92.67	2.42	0.497	21%	42%
CampfireFHD10_0_3	2	2	2	2	2	2	3	0.96	65.19	2.20	0.403	18%	20%
CampfireFHD10_0_5	2	2	2	2	2	2	2	0.95	31.40	2.00	0.451	23%	10%
CampfireFHD10_0_75	2	2	2	2	1	1	2	0.92	30.37	1.70	0.462	27%	0%
CampfireFHD10_1	1	2	1	1	1	2	1	0.90	30.04	1.37	0.486	36%	0%

Fig. 2. Example of data presented in the spreadsheet.

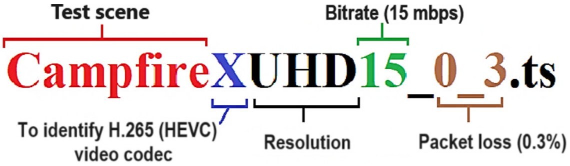
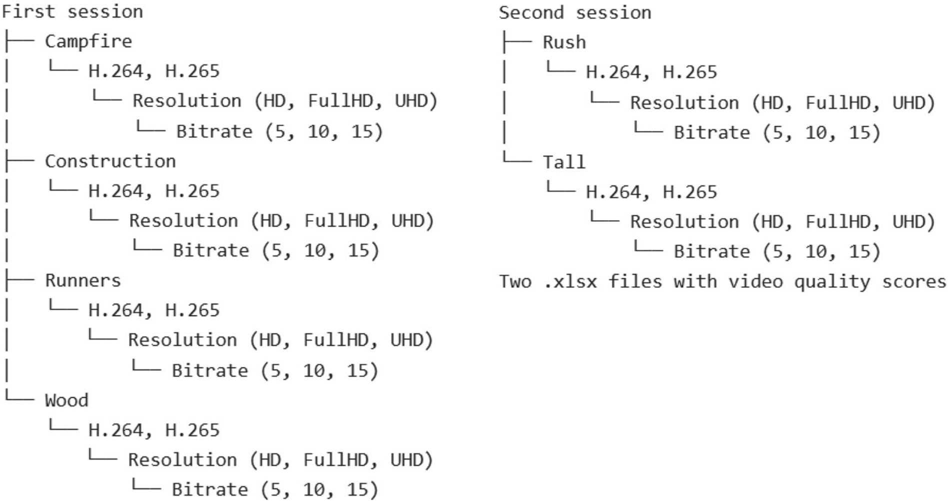


Fig. 3. Example of the naming pattern.



```
First session
├── Campfire
│   └── H.264, H.265
│       └── Resolution (HD, FullHD, UHD)
│           └── Bitrate (5, 10, 15)
├── Construction
│   └── H.264, H.265
│       └── Resolution (HD, FullHD, UHD)
│           └── Bitrate (5, 10, 15)
├── Runners
│   └── H.264, H.265
│       └── Resolution (HD, FullHD, UHD)
│           └── Bitrate (5, 10, 15)
└── Wood
    └── H.264, H.265
        └── Resolution (HD, FullHD, UHD)
            └── Bitrate (5, 10, 15)

Second session
├── Rush
│   └── H.264, H.265
│       └── Resolution (HD, FullHD, UHD)
│           └── Bitrate (5, 10, 15)
└── Tall
    └── H.264, H.265
        └── Resolution (HD, FullHD, UHD)
            └── Bitrate (5, 10, 15)
Two .xlsx files with video quality scores
```

Fig. 4. Structure of the dataset.

higher than 50 %, dispersion around the mean is too high and the mean should not be used as a parameter describing the selected population. Variable Fair or Better (FoB) represents the ratio of satisfied users (MOS => 3) [11]. Part of the spreadsheet is shown in Fig. 2. The method used for naming video sequences is depicted in Fig. 3. The logical structure of directories is shown in Fig. 4.

4. Experimental Design, Materials and Methods

For the purpose of the experiments, we created a video streaming testbed. The testbed consisted of a video streaming server represented by the FFmpeg tool and a PC with VLC player as end terminal. We ensured that local network packets were dropped only on the network interface card installed on the PC and were not affected by any unwanted phenomenon that could have occurred for real data traffic in a packet-based network. Packets were randomly discarded by the Clumsy software that allowed us to set a packet drop ratio. Media streaming was based on the RTP/UDP protocols. RTP or Real-time Transport Protocol is a protocol

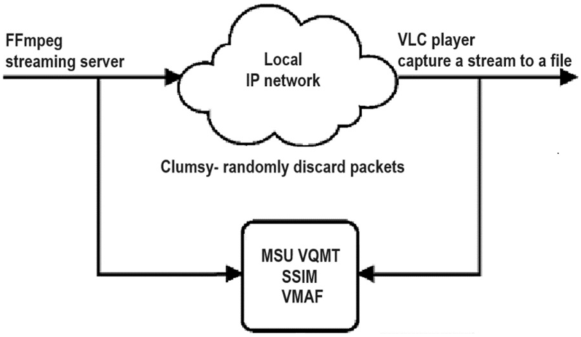


Fig. 5. Video streaming testbed.

for streaming multimedia content over the Internet. UDP or User Datagram Protocol is a communication protocol used across the Internet especially for time-sensitive transmissions (e.g. voice or video streaming). It speeds up communication because it relies on a connectionless communication model with a no handshake mechanism related to the retransmission of undelivered packets. The scheme of our implemented testbed is depicted in Fig. 5.

The FFmpeg tool was also used to prepare the test video sequences. An example command for encoding is:

*ffmpeg -f rawvideo -video_size 3840 × 2160 -pix_fmt yuv420p -framerate 30 -i C:\Users:\ Runners.yuv -vcodec libx264 -x264-params keyint=15:*min-*keyint=15:bframes=3:b-adapt=1:bitrate=5000:vbv-maxrate=5000:vbv-bufsize=5000 -vf scale=3840 × 2160 RunnersUHD5.ts* where:

- libx264/libx265 – video encoder,
- framerate 30 frames per second (300 frames total = video length is 10 s),
- bitrate - value in kbps,
- vfscale – resolution of the encoded video file.

Video was streamed by the following command:

ffmpeg -re -i source video file address -c copy -f mpegts udp://IP address:1234 and captured by the VLC player:

udp://@IP address:1234

Each video file was rated by real observers. We organized two subjective assessment rounds. In the first round, there was 60 participants (Women:22, Men:38; age group: 18–35), and there was 46 in the second round (Women: 32, Men:14; same age group). According to the ITU recommendation [6], the minimal number of participants for a session in a laboratory environment should be at least 15. The experimental sessions met all the necessary requirements, including breaks, equipment, viewing distance, and familiarization with the voting procedure and pace. The subjective ACR method is a category judgment. Test sequences are presented one after the other. After each presentation, observers are asked to evaluate the quality of the sequence shown on a MOS scale. While ACR does not require a reference video, the two objective methods SSIM and VMAF are full reference metrics. In other words, the degree of deterioration is measured as the difference between the picture quality of the reference and the test sample. The calculation process of SSIM is derived from an analysis of contrast, brightness, and structural similarity with the reference video sequence. The VMAF metric considers information fidelity loss, as well as measures loss of details and impairments that distract viewer. Both metrics were calculated by MSU VQMT 14 Pro-version.

Limitations

A few test sequences reached a CoV higher than 50 % (19 video sequences, 2 %), which indicates that the mean cannot be used to represent such video sequences. These sequences were not included in the spreadsheet and dataset.

Ethics Statement

The dataset collected in this study did not involve animals, and it did not contain data collected from social media platforms either. Informed consent was obtained from respondents before the start of the subjective assessment session.

Data Availability

Video dataset for coding efficiency and image quality analysis containing results from standardized objective and subjective assessment methods. (Original data) (Mendeley Data).

CRediT Author Statement

Jaroslav Frnda: Conceptualization, Methodology, Investigation, Writing – original draft; **Marek Durica:** Formal analysis, Data curation; **Jerry Chun-Wei Lin:** Visualization, Supervision, Writing – review & editing; **Philippe Fournier-Viger:** Supervision, Validation, Writing – review & editing.

Acknowledgments

This research did not receive any specific grant from funding agencies in the public, commercial, or not-for-profit sectors.

Declaration of Competing Interest

The authors declare no competing interest.

References

[1] Recommendation ITU-T P.1204: video quality assessment of streaming services over reliable transport for resolutions up to 4K, Switzerland 2023.

[2] M. García-Torres, D.P. Pinto-Roa, C. Núñez-Castillo, B. Quiñonez, G. Vázquez, M. Allegretti, et al., Feature selection applied to QoS/QoE modeling on video and web-based mobile data services: an ordinal approach, Comput. Commun. 217 (2024) 230– 245.

[3] T. Hoßfeld, P.E. Heegaard, M. Varela, L. Skorin-Kapov, M. Fiedler, From QoS distributions to QoE distributions: a system's perspective, in: 2020 6th IEEE Conference on Network Softwarization (NetSoft), Ghent, Belgium, 2020, pp. 51–56.

[4] C.G. Bampis, Z. Li, I. Katsavounidis, T.-Y. Huang, C. Ekanadham, A.C. Bovik, Towards perceptually optimized adaptive video streaming-a realistic quality of experience database, IEEE Trans. Image Process. 30 (2021) 5182–5197.

[5] A. Mercat, A. Makinen, J. Sainio, A. Lemmetti, M. Viitanen, J. Vanne, Comparative rate-distortion-complexity analysis of VVC and HEVC Video codecs, IEEE Access 9 (2021) 67813–67828.

[6] Recommendation ITU-T P.910: subjective video quality assessment methods for multimedia applications, Switzerland 2022.

[7] L. Song, X. Tang, W. Zhang, X. Yang, P. Xia, The SJTU 4K video sequence dataset, the Fifth International workshop on quality of multimedia experience (QoMEX2013), Klagenfurt, Austria, July 3rd–5th, 2013.

[8] Zhou Wang, A.C. Bovik, H.R. Sheikh and E.P. Simoncelli. Image quality assessment: from error visibility to structural similarity, in IEEE Trans. Image Process., vol. 13, no. 4, pp. 600–612, 2004, doi: 10.1109/TIP.2003.819861.

[9] VMAF - video multi-method assessment fusion. Netflix. Available on Github: https://github.com/Netflix/vmaf.

[10] Recommendation ITU-R BT.500-15. Methodologies for the subjective assessment of the quality of television images, Switzerland 2023.

[11] T. Hoßfeld, P.E. Heegaard, M. Varela, et al., QoE beyond the MOS: an in-depth look at QoE via better metrics and their relation to MOS, Qual. User Exp. 1 (2) (2016), doi:10.1007/s41233-016-0002-1.