# Objective Models for Performance Comparison of Compression Algorithms for 3DTV

*Jan KUFA, Ondrej KALLER, Ondrej ZACH, Ladislav POLAK, Tomas KRATOCHVIL*

Dept. of Radio Electronics, Brno University of Technology, Technicka 12, 616 00 Brno, Czech Republic

{xkalle00, xzacho04}@stud.feec.vutbr.cz, {kufa, polakl, kratot}@feec.vutbr.cz

Submitted May 10, 2018 / Accepted November 6, 2018

**Abstract.** *Efficient video compression algorithms in advanced multimedia broadcasting systems are in high demand. In the last decades, different video compression tools have been developed which can influence the final Quality of Experience in different ways. his paper has two goals. The first goal is to present a study of different compression algorithms available for stereoscopic 3D videos. The second goal is to present the possibilities in the creation of new stereoscopic models. The well-established video codecs (AVC, MVC, HEVC and MV-HEVC) are considered as encoders. Generic objective video quality metrics are used to analyze the compression efficiencies of the considered codecs, extended with results from subjective tests. The correlations between the objective and subjective scores are analyzed statistically. Due to unsatisfactory results of generic 2D metrics for the stereoscopic sequences used in the test, new objective models are presented. Such models show improved correlation with subjective stereoscopic video quality. The validation, verification and a description of models are presented in detail.*

## Keywords

3D video coding, AVC, MVC, HEVC, MV-HEVC, objective and subjective video quality metrics, MOS, Spearman and Pearson rank correlation

## 1. Introduction

Nowadays, interest in excellent video quality is rapidly increasing. Such interest is closely related to the provided video services in Standard and High Definition (SD and HD) formats and in the future in Ultra HD (UHD) or Three-dimensions (3D). UHD and 3D are emerging video formats with specific features. It is evident that flexible and highly efficient video coding algorithms are very important to distribute video content in such formats and in a required quality [1]. As an example, we can state the scenario where we are strongly limited by transmission data rate, which is often the case in wireless networks. Furthermore, we may be limited in the maximum amount of transferred data, so-called Fair User Policy (FUP), in mobile networks [2].

Today's display units already reach technical properties that are close to the resolution limits of the human eye. Enhanced display qualities, such as Ultra HD, High Dynamic Range (HDR), High Frame Rate (HFR), and wide color gamut are already approaching the limit of Human Visual System (HVS) in terms of user experience for 2D video. An appropriate approach to represent a real 3D view can be one of the next research directions. For instance, holographic displays and volumetric displays.

Video compression tools play a key role in the fulfillment of both multimedia content provider's requirements (e.g. bandwidth needed for transmission) and users' requirements (transparent video quality). When using any compression tool, it is important to find a balance between the compression ratio and user's Quality of Experience (QoE). A high compression ratio can significantly reduce the amount of data in the processed video but results in high degradation of video quality. The assessment of such a degradation is especially important for 3D visual content which has been receiving attention in many fields of interest (e.g. TV broadcasting, security, medicine). Consequently, accurate evaluation of stereoscopic 3D video quality by objective and subjective metrics is highly required [3]. Despite the fact that today's interest of stereoscopic television seems to be out of date, new publications that deal with this issue keep appearing – as an example, we can refer to publication [4].

The paper aim is to explore the performance of recent and emerging compression tools for 3D stereoscopic video, namely H.264 Advanced Video Coding (AVC) [5], H.264 Annex H - Multiview Video Coding (MVC), H.265 High Efficiency Video Coding (HEVC) [6] and H.265 Annex G - Multiview High Efficiency Video Coding (MV-HEVC) [7]. For this purpose, appropriate subjective test sessions have been realized. Well established 2D objective video quality metrics are then compared with scores from the subjective test. Moreover, gathered results are statistically analysed. Based on subjective test results and commonly used 2D metrics, models of 3D stereoscopic metrics are developed to best describe the quality of stereoscopic videos. Our general development of objective models can also be applied to non-3D video types, such as UHD, etc. The results presented in this paper are a continuation of our earlier work published in [8].

The rest of this paper is organized as follows. The related state-of-the-art and the main contributions of this research paper are described in Sec. 2. The test setup is described in Sec. 3, including the used subjective video quality method, its setup and the whole realization. Section 4 contains the results of the objective metrics and subjective test and their further evaluation and discussion. Section 5 describes the proposal and verification of our models for stereoscopic video. Finally, conclusion remarks are outlined in Sec. 6.

# 2. Related Work

There are several possibilities how to encode stereoscopic video content. Each view of the stereo pair can be encoded separately as an independent video sequence using common video coding algorithms for 2D video sequences. Another possibility is to use video coding algorithms specifically designed to support multiple views. These algorithms usually consider the similarity of both views which can lead to significant bitrate savings. Also, specialized video coding algorithms for 3D exist which can take advantage of depth maps if present. The following paragraphs relate to previously published works regarding video coding of 3D content for multimedia purposes and related Quality of Experience.

In recent years, numerous studies have focused on exploring the possibilities of encoding stereoscopic 3D video content. Hannuksela et al. [9] offer an extensive overview of the MultiView extension of the High Efficiency Video Coding standard. MV-HEVC is capable of encoding multiple views together without using a depth map and is also able to encode stereoscopic 3D video. The overview of the 3D extension of HEVC (3D-HEVC) is presented in [10]. As 3D-HEVC is designed for encoding 3D content, it utilizes both the stereo pair and the information from the depth map and camera configuration. Results of software evaluations suggest that it is possible to achieve about 52 % coding efficiency gain on average when using 3D-HEVC compared to standard MVC. A special case is described in [11], where an extension of 3D-HEVC considering circular camera arrangement is proposed.

## 2.1 Objective Metrics and Models in Stereoscopic 3DTV

Possibilities of using common 2D objective metrics for stereoscopic video were examined in [12] and [13]. In the first paper, the impact of encoding artefacts in stereoscopic video quality has been evaluated with three 2D objective metrics. The evaluation was done using Peak Signal-to-Noise Ratio (PSNR), Structural Similarity (SSIM) and Visual Information Fidelity- pixel domain (VIFp). The results show that only the VIFp results were highly correlated with subjective data among selected objective metrics. The paper [14] investigates the reliability of objective quality metrics commonly used for the quality assessment of 2D media, in the context of stereoscopic 3D Video. The consistency between objective and subjective measures is evaluated by the Pearson linear correlation coefficients (PLCC). In [15], the use of 2D objective metrics for 3D quality assessment has been explored. Two objective metrics, Video Quality Metric (VQM) and Perceptual Quality Metric (PQM), have been investigated and their alignment to the Mean Opinion Score (MOS) has been analyzed. In that paper, unlike ours, the video sequences were encoded only by using the AVC encoder. Based on the statistical Pearson correlation (PCC) analysis, PQM correlates better with MOS than VQM, 0.78 versus 0.97. Results also indicated that the correlation is strongly content dependent. In another work, Han et al. [16] proposed an extended no-reference objective stereoscopic 3D Video Quality Metric (eNVQM) for 3D video quality assessment. Performance of eNVQM was studied in comparison with two 2D objective video quality metrics, SSIM and VQM. The PCC analysis showed that eNVQM has better accuracy, PCC equal to 0.944, in terms of human perception for stereoscopic video, compared to two current common assessment methods. Pearson correlation for SSIM was 0.911 and 0.932 for VQM.

## 2.2 Subjective Assessment in Stereoscopic 3DTV

The authors of [17] analyzed the possible use of Absolute Category Rating (ACR) for 3D stereoscopic content. A study of subjective quality of monoscopic and stereoscopic video in adaptive streaming in [18] presents a comparative analysis of different bitrate adaptation strategies in adaptive streaming in 2D and 3D scenarios. We can observe that if the experiment was done on monoscopic video content then no statistical differences were found when changing the bitrate in an abrupt or a gradual way. Also, high quality oscillations were hardly perceptible if there is not so much coding bitrate difference. Tests on stereoscopic video confirms that switching from 3D to 2D could be the best option to reduce the bitrate, while the inverse behavior does not provide a significant improvement to QoE. Paper [19] studied the response of the HVS to compressed stereoscopic sequences and compared the visibility of artifacts in 3D and 2D views (individually left and right eye views) over a different range of bitrates. The 2D and 3D MOS from the test showed that there is a bitrate threshold above which compression artifacts tend to be suppressed in the 3D view when compared to the classic 2D view. The correlation between objective metrics and subjective tests is highly depending on the features of the used video sequences. It is therefore appropriate to perform extensive tests with different methods (subjective), codecs and videos.

Based on the brief state-of-the-art presented above, our paper tries to answer the following points:

1) Which 2D objective metrics correlate best with the users Mean Opinion Score (MOS) for stereoscopic 3D videos, based on additional statistical analyzes?

2) Can any 2D objective metrics be optimized for better 3D accuracy? Alternatively, can a model be created that has a better correlation with stereoscopic 3D video sequences?

# 3. Experimental Setup

This section briefly describes the setup of our experiment. The video sequences used for the test, parameters of encoders and used objective metrics for rating the quality of video sequences are outlined.

## 3.1 Video Sequences

As the source of stereoscopic 3D video sequences, we have used four samples which are available in databases [20] and [21], to make our research have a wider range of uses. hese videos are used as an original dataset for our test. The additional four video sequences were used at another subjective test for the verification of the proposed objective models [21]. All these sequences were in Full HD resolution (1080p) for each view and had a frame rate of 25 frames per second (fps). The length of each video sequence before encoding was adjusted to 10 seconds, which is a typical length used in subjective video quality studies. The selected video sequences cover a wide variety of contents as can be seen from the Spatial Information (SI) and Temporal Information (TI) in Fig. 1. The figure also contains one frame of each corresponding sequence. Both parameters SI and TI were calculated according to ITU-T P.910 [22]. The average value of depth for 5, 50 and 95 % for each video sequence was calculated, as can be seen in Tab. 1. According to the obtained results, for instance, in the case of video Train, 95 % of pixels will have a depth (shift of pixels) of 15.27. In other case, video Basketball will have 5 % of the pixels with a depth of −10.59. For the calculation of the average value of depth, software StereoPhoto Maker was used[1]. The average value of the depth of the videos varies considerably.

| **Image depth** [%] | **Videosequence** | | | |
|---|---|---|---|---|
| | ***Basketball*** | ***Poznan Hall*** | ***Train*** | ***Wishing Well*** |
| **Depth d05** | −10.59 | −40.95 | 8.17 | −24.06 |
| **Depth d50** | −3.96 | −24.25 | 12.40 | −11.63 |
| **Depth d95** | 3.66 | −12.54 | 15.27 | 18.56 |

**Tab. 1.** Average depth of video sequences.

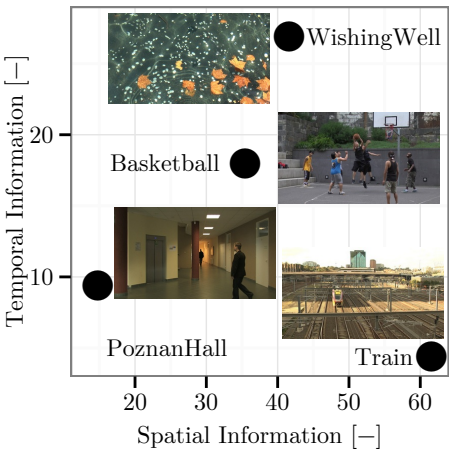



**Fig. 1.** Spatial and temporal indexes of used sequences.

[1]www.stereo.jpn.org/eng/stphmkr

The content of the video sequences can be described as follows:

- Basketball: Basketball players playing in the street. A moving camera with a wide shot. Fast and unpredictable movements. Different physics including jumping players and ball in the air.
- PoznanHall: A view into a school corridor with a slightly moving camera and a walking man in the foreground. Slow and predictable movements.
- Train: Static view of a train station with an approaching train with a detailed view of the overhead wire. Slow motion video and predictable movement of the train.
- WishingWell: Fountain with coins on the bottom and moving leaves on the surface. A lot of small waves on the water and reflections from the water.

## 3.2 Encoding Parameters

As input to encoders, only bitrates were defined together with searching motion range without any other system parameter modification and without any tuning of the encoders. The quality profile was set to the highest quality because we were focusing only on the quality of encoding, not the encoding speed. Encoders selected other parameters automatically by itself. A summary of video encoders settings used in encoding is provided in Tab. 2. Target bitrates were adjusted between 0.5 Mbps and 4 Mbps. The target bitrate applies to one view only. Let us give an example for the 1 Mbps bitrate: For the 2D encoder, the total bitrate 1 + 1 Mbps was set. For multiview encoders, the stream of both views was 2 Mbps. This means that the total data rate is the same. The searching range for HEVC-based encoders was set to 64 pixels to take full advantage of these modern encoders.

## 3.3 Objective Video Quality Metrics

A specific feature of stereoscopic 3D videos is a broad variety of imaging technologies available. They have a different structure of image data and different types of compression standards. Currently, there is no widespread general objective 3D metric. Therefore, widely established general metrics like PSNR, SSIM, and VQM are commonly used. Also more advanced metrics for exist a particular type of content or compression that achieve better results for their specific area of use. However, the intention of this contribution is to create a general method for stereoscopic content. The PSNR is a very simple metric based on differences of the corresponding pixel values [23]. The value of PSNR was computed for the luminance component only. The SSIM computes the structural differences in the pictures reflecting basic properties of the HVS [23]. Finally, VQM compares the original characteristics with the processed characteristics of the video sequences and then it produces VQM scores. The range can be from 0 (no perceived deterioration) to approximately 1 (maximum perceived deterioration). A general model was used for our case [24].

| Codec | AVC | MVC | HEVC | MV-HEVC |
|---|---|---|---|---|
| **Implementation** | x264 r2597 | FRIM x64 1.25 | HM 15 | HTM 15.1 |
| **Profile** | High | High | Main | Main |
| **Level** | 5.1 | 4.0 | 5.1 | None |
| **Preset** | Very Slow | 1 (quality) | / | / |
| **Search Range** | 32 | 32 | 64 | 64 |
| **GOP Size** | 8 | 25 | 8 | 8 |
| **Entropy coding** | CABAC | CABAC | CABAC | CABAC |

**Tab. 2.** Parameters of used encoders.

### 3.4 Subjective Test Setup

All subjective video quality assessment was conducted in a special test room. Laboratory conditions were set up according to ITU-R BT.500-13 [25] including a room with controlled lighting. For the subjective experiment, a plasma stereoscopic television (Panasonic TX-P42GT20E) was used to display stereoscopic 3D video content. The television's active shutter 3D system with a Full HD double frame rate was used. In contrast to polarization 3D system, this method of 3D view does not reduce the resolution. It is its biggest advantage. The video format structure of the Frame Packing 3D was used. It conveys to two "full resolution" 1080p video signals, one for each eye, to the TV. This method is marked as Full High Definition 3D (FHD3D). The interface used between TV and PC was HDMI 1.4 which is capable to successfully transfer FHD3D. The peak luminance of the display was adjusted to 200 cd/m$^2$. The viewing distance of the participants from the display, according to ITU-R BT.2022 [26], is the height of the picture multiplied by 3.2. In our case, the optimal viewing distance is 1.7 meters (see Fig. 2). In the subjective test, only one participant was in front of the television to eliminate the effect of different observation positions.

As the pretest 3D sequences in three qualities were played. These sequence were different from the sequences used during the test. Observers had an overview of how the 3D movie could look. Sequences were randomly played for each participant who did not know about the details (the used encoder or bit rate). The observers were asked for evaluating the quality of the played video using a simple five point discrete scale in the range from 1 (Bad) to 5 (Excellent). Whole tested sequences were evaluated by all participants for the best consistency of results. Participants rated the quality on sliders which were connected to the master computer (see Fig. 2). This computer also controls the media computer from which the video sequences were played.

### 3.5 Participants and Color Vision Test

A total group of 37 observers participated in the stereoscopic 3D subjective test. Two of them were female. The youngest participant was twenty years old and the oldest one was forty-two years old. Overall, eight of them had some experience with video quality assessment. Next, five observes previously participated in stereoscopic 3D video quality assessment tests. University students and employees were recruited with an average age of 24.

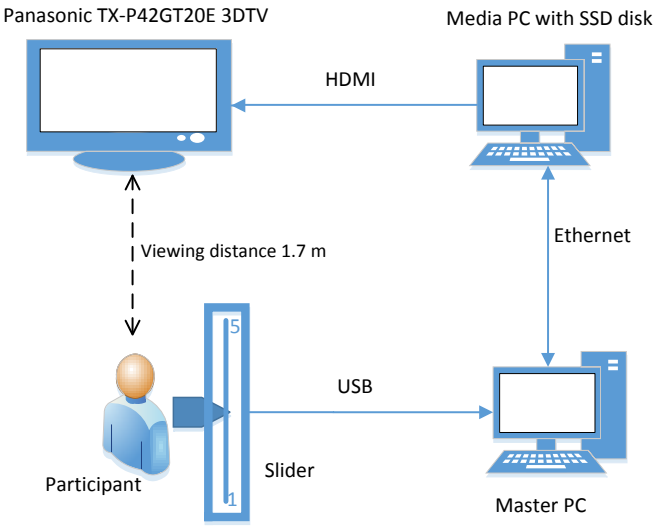



**Fig. 2.** Block diagram of the subjective test setup.

The youngest participant was 20 years old and the oldest was 42 years old. Color blindness (Ishihara test) of all participants was tested as well as their ability of stereoscopic vision (Randot stereotest) [27]. Three people who did not pass the tests were not included in the final evaluation.

## 4. Experimental Results

The results obtained from the objective metrics, subjective test and additional correlation tests are evaluated, compared and discussed in this section.

### 4.1 Coding Efficiency According to the Objective Metrics

The results show that the performance of standard codec and mutual comparison is highly content dependent. For several content types, the multiview coding gain is significant, while for other contents the multiview coding only brings undesired overhead with no performance improvement. Overall, the codecs belonging to the same standard exhibit very similar performance with differences in PSNR in the order of 1 to 3 dB. Results obtained from objective metrics and subjective test are shown in Fig. 3. The MOS has been evaluated together with the 95% confidence interval. A detailed analysis of the results is published in [8].

## 4.2 Coding Efficiency Evaluation Based on the Subjective Test

Advanced results and analysis of subjective tests for all sequences and codecs are presented in Fig. 4. Legend in the figure is presented as follows: "First is the numbering and after that is the name of the video sequence, used encoder and target bitrate. The last columns presents the MOS". For example, in the first row, the second line is the sequence "Basketball" encoded by AVC with bitrate 1 Mbps. The central red mark in MOS is the median, the edges of the blue box are the 25th and 75th percentiles. The most extreme data points, without outliers, are the black whiskers. Outliers (Red Cross) are plotted individually. The following lines describe the subjective test results in Fig. 4:

Basketball: The performance of AVC and MVC is very similar. In addition to the highest bit rate, there MVC is better. For HEVC and MV-HEVC, the quality is the same for all bitrates. There is no increase in quality between the bit rates 2 and 4 Mbps. The results of the subjective tests correspond approximately to the objective metrics.

PoznanHall: In the case of the HEVC codec, there is a gradual increase in quality with higher bit rates. On the other hand, with MV-HEVC, the quality was similar for all bit rates.

Train: The coding efficiency of HEVC and MV-HEVC is similar. Bit rate higher than 1 Mbps does not cause predicted improvement in the QoE. In the case where bit rate is higher than 2 Mbps, then the coding efficiency is similar for all codecs. There is no coding gain of the multiview codecs.

WishingWell: The performance of MVC is significantly better than AVC. It is a situation in which the codec is able to exploit multiview coding potential. Similar results were obtained for codecs HEVC and MV-HEVC.

The results of the Wilcoxon signed-rank test [28] are presented in Tab. 3. This test did not reject the hypothesis that AVC and MVC have similar coding efficiency. The same result is also obtained for HEVC and MV-HEVC. The H.265 standard was designed to produce a 50 % less bitrate compared to H.264 standard for the same image quality [6]. The hypothesis that H.265 generation needs half the bitrate to compare to H.264, for the same quality also in stereoscopic video, has been proved ($p$ equal to 0.31). If the value of $p$ would be very small then the hypothesis would have not been proved (for example, number 0.02).

| Hypothesis about MOS | Rejection of the hypothesis | $p$ [-] |
|---|---|---|
| AVC ≈ MVC | 0 | 0.6483 |
| HEVC ≈ MV-HEVC | 0 | 0.4616 |
| Codecs H.264 vs. H.265 compression efficiency in stereoscopic video | | |
| H.265 has double compress efficiency | 0 | 0.3128 |

**Tab. 3.** Results of Wilcoxon signed-rank test.

The results show that the scattering of the test subject's evaluation in the subjective test is large [29], [30]. For this reason, it was necessary to evaluate the participants who acted as outliers. We have used the whisker method for outlier values detection. Whisker (w) extends the interval of quartiles (Q25, Q75) by $w$ on both sides. The whiskers are lines extending above and below each box (Q25–Q75). Whiskers are drawn from the ends of the interquartile ranges to the furthest observations within the whisker length. Observations beyond the whisker length are marked as outliers. In our case of normal distribution, w is equal to 1.5, which would correspond 99.3 percentiles coverage of values. An outlier is a value that is more than 1.5 times the interquartile range away from the top or bottom of the box [31].

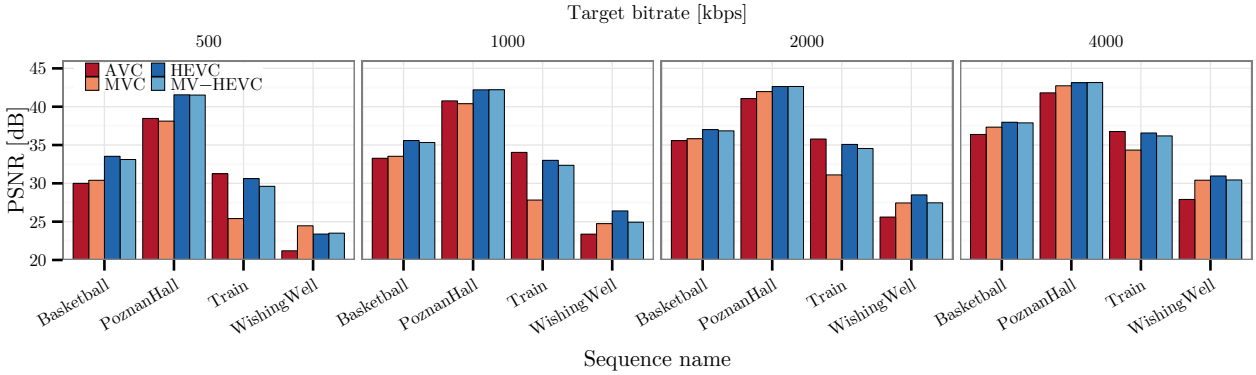
There are two hypotheses about outlier rate. First, there is a difference in variance of QoE evaluation among the sequences. Second, the variance is larger at the beginning and at the end of the testing session, due to disorientation and fatigue. A hypothesis was tested concerning uniform distribution of outliers through sequences and time.

The Hi-square goodness-of-fit tests against discrete uniform distribution, in the case of sequences and time, have rejected this proposition at a significance level of 0.05 [32]. We can prove that in our subjective test, after significantly lower outlier parts (first 8 video sequences) the rest of the evaluation time has uniform outlier rate. This hypothesis can be seen in Figs. 5 and 6. In these figures, the blue color indicates the results which were below the permissible deviation. The results marked in yellow color are those that were above the error of the mean. The data from participants, which has more than 10 % of outlier evaluations, was excluded. The number of participants not included in the final evaluation is four, which amounts to 11.8% of the test base.

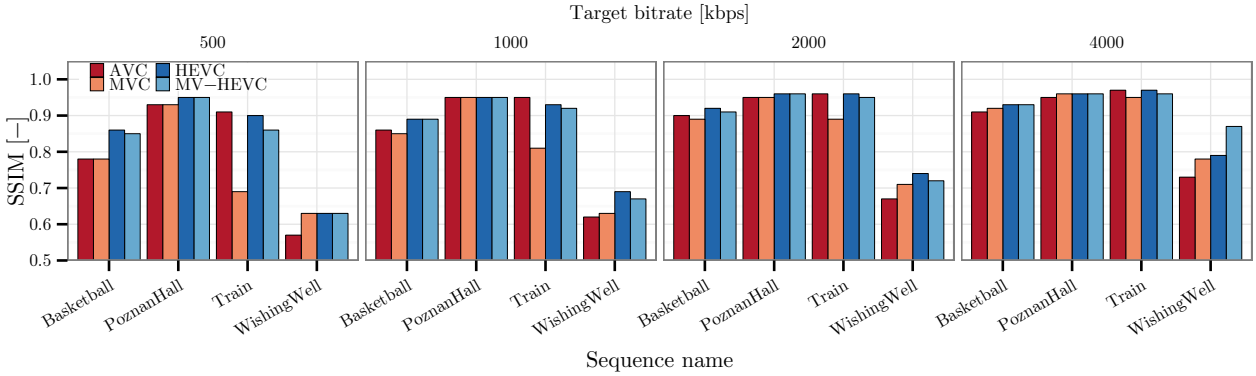
## 4.3 Correlation of Subjective and Objective Metrics for Stereoscopic 3D Video

After evaluating the coding efficiency, it is necessary to determine which 2D metric has the greatest correlation with the subjective 3D test. For these purposes, Spearman's Rank Order Correlation Coefficient (SROCC) and PCC were applied to the results [33]. Other evaluation methods of models performance, with respect to subjective tests, for objective quality assessment are described in [29]. These analyses are used to determine the correlation between objective and subjective metrics.

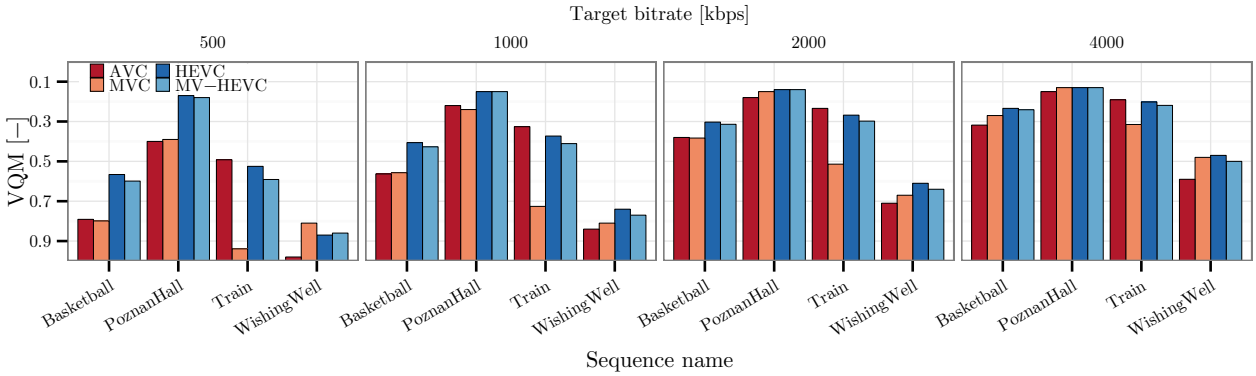
The correlation scores are between +1 and −1, where −1 and +1 mean total positive and negative linear correlation respectively, and 0 denotes no linear correlation. Next possible method for evaluation of results is based on ROC curves [34]. The VQM objective metrics has negative values in correlation, because their lower score indicates higher video quality.



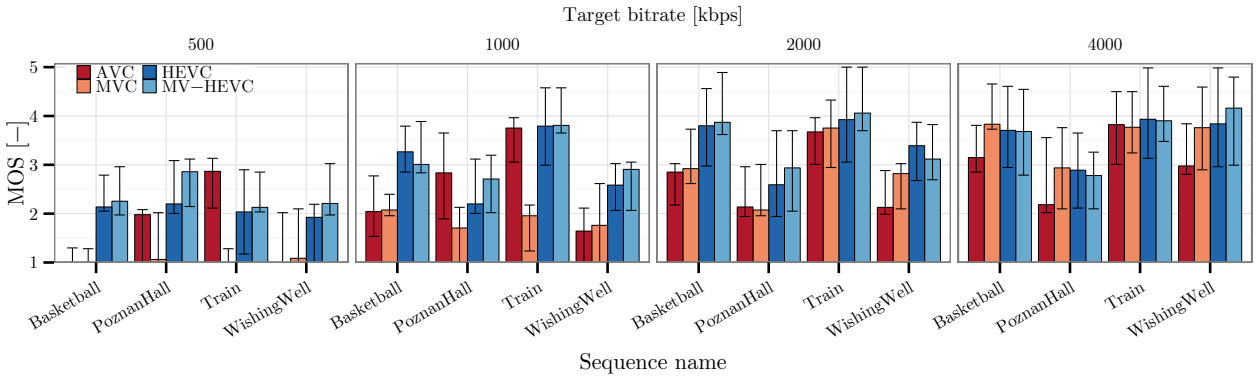

**(a)** Results of PSNR metrics.




**(b)** Results of SSIM metrics.




**(c)** Results of VQM metrics.




**(d)** Results of subjective test with 95 % confidence interval.

**Fig. 3.** Video quality measured by objective metrics and subjective test.

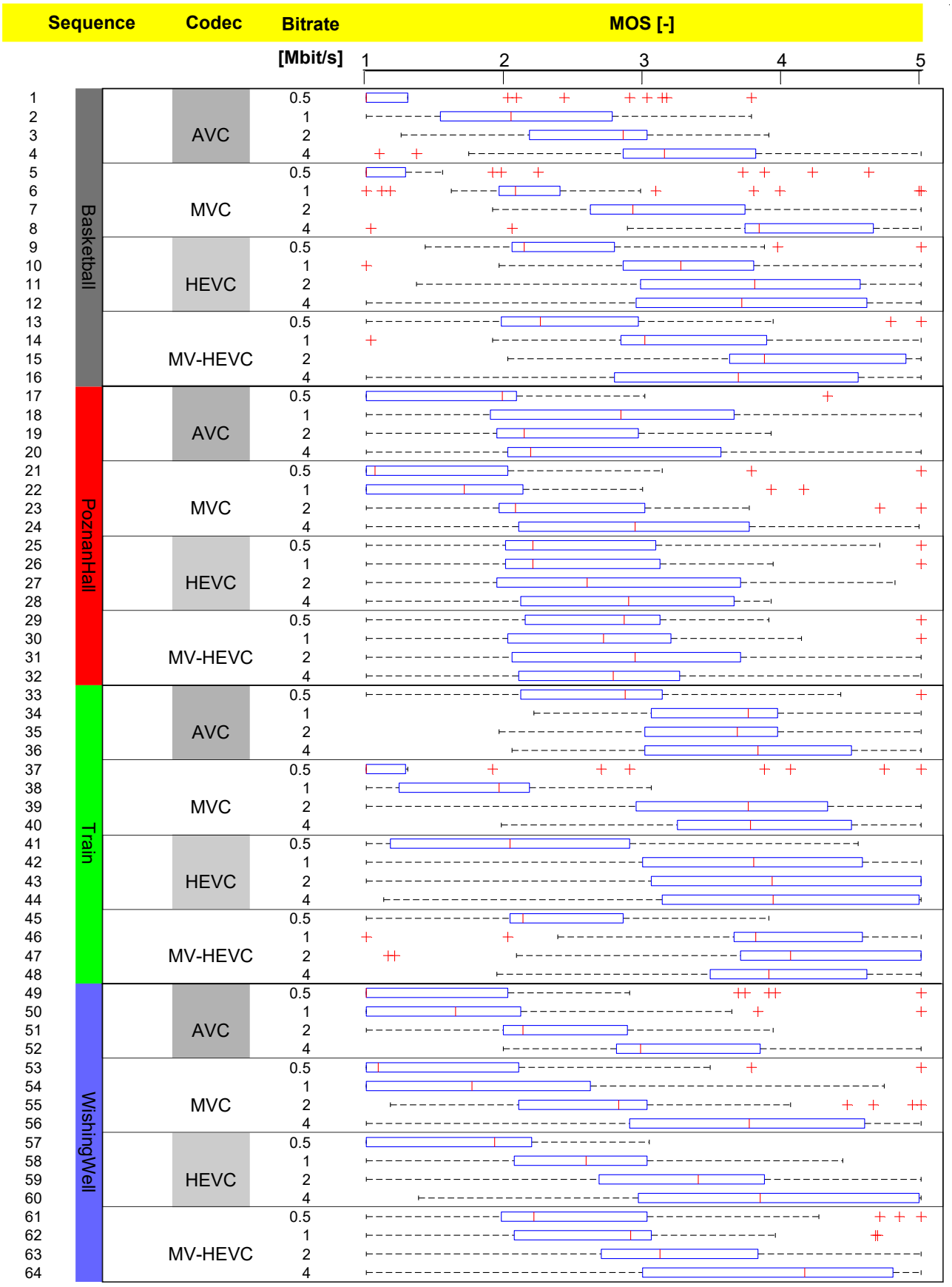



**Fig. 4.** Results of the subjective test in detail.

| Video sequence | PSNR | | SSIM | | VQM | |
|---|---|---|---|---|---|---|
| | *Pearson* | *Spearman* | *Pearson* | *Spearman* | *Pearson* | *Spearman* |
| **Basketball** | 0.980 | 0.919 | 0.965 | 0.899 | −0.974 | −0.892 |
| **PoznanHall** | 0.261 | 0.442 | 0.293 | 0.472 | −0.502 | −0.484 |
| **Train** | 0.606 | 0.648 | 0.848 | 0.696 | −0.802 | −0.648 |
| **WishingWell** | 0.872 | 0.939 | 0.888 | 0.965 | −0.877 | −0.905 |
| $\sum$ | 0.261 | 0.226 | 0.462 | 0.415 | −0.511 | −0.407 |
| **Seq.- 1.,3.,4.** | 0.608 | 0.625 | 0.672 | 0.702 | −0.815 | −0.792 |

**Tab. 4.** Spearman's rank order correlation coefficient and Pearson correlation coefficient.

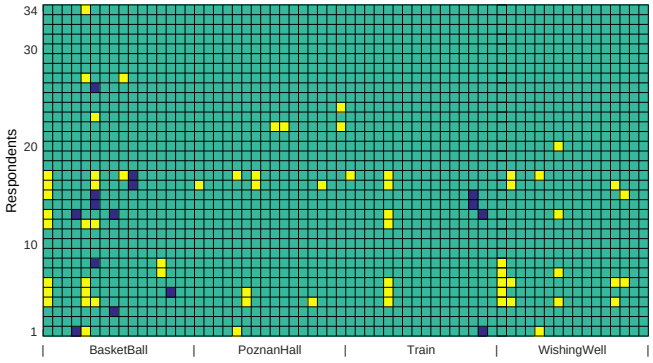



**Fig. 5.** Dependence of the number of outliers on the sequence.

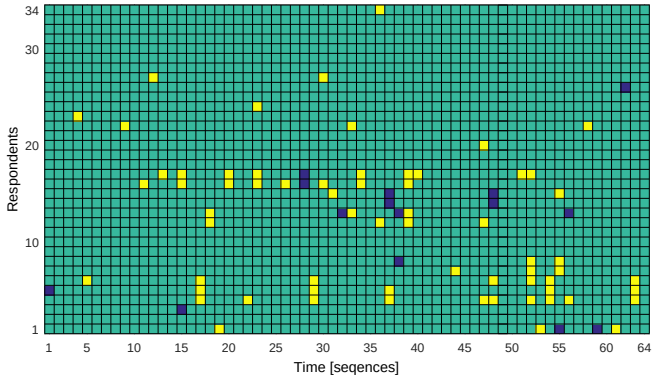
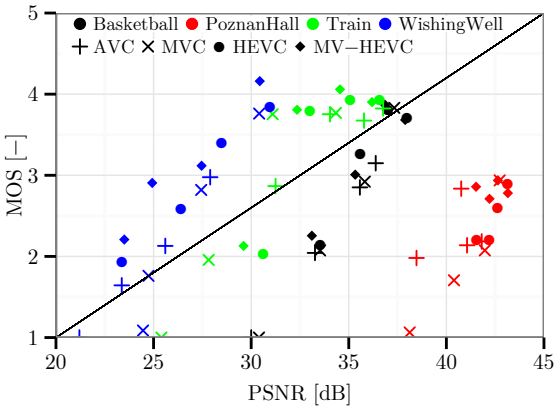


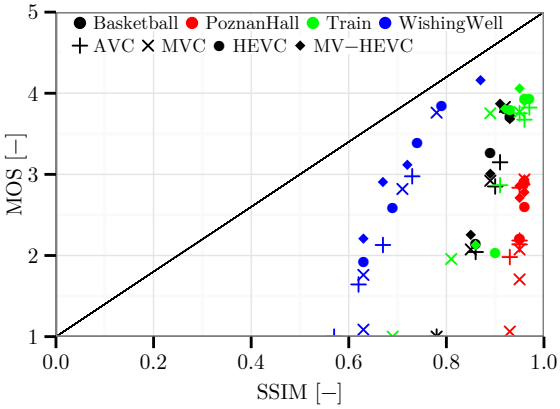

**Fig. 6.** Number of outliers depending on playing time.

The correlation analysis was firstly applied to each sequence separately (see Tab. 4). Due to the fact that we need a universal metric, the correlation value was then calculated across all sequences. The results show that the correlation depends on the video content. The video "PoznanHall", which is from another video database, has a different correlation than other videos. It may also be due to the fact that the video has a large stereoscopic parallax (see Tab. 1). For some viewers, it could be distracting and therefore the video has a non-standard rating. For this reason, in the last row of the Tab. 4, the "PoznanHall" sequence is omitted and the resulting score, just in this row, is calculated without it.

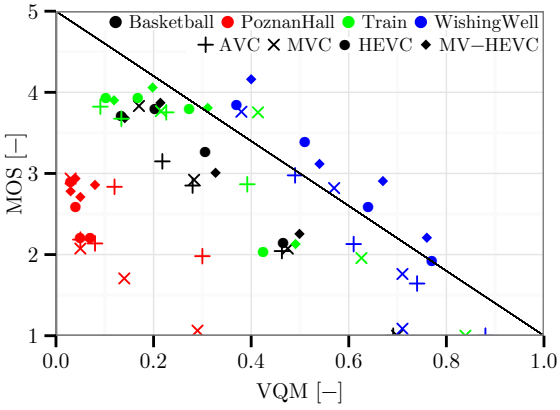
The correlation between objective and subjective methods is plotted in Fig. 7. The black markers represent the video "Basketball", whereas red, green and blue colors indicate videos "PoznanHall", "Train" and "WishingWell", respectively. After a thorough comparison of all objective and subjective scores, it can be concluded that in our case the VQM objective metric best reflects the user's QoE for compressed stereoscopic 3D videos.




**(a)** Correlation between the PSNR metric and the subjective test.




**(b)** Correlation between the SSIM metric and the subjective test.




**(c)** Correlation between the VQM metric and the subjective test.

**Fig. 7.** Dependence of objective metrics on the subjective test.

# 5. Innovative Models for Stereoscopic 3D Video Content

Although the VQM metric has the highest correlation, it is still not ideal for evaluating stereoscopic 3D videos. We thus propose our own model, which better models our subjective test results. Such a model should provide sufficient general predictions at least for content with similar parameters as the used video sequences. This section describes the model proposal, validation and verification of the models, and a description of each proposed model is provided at the end.

We have several objective parameters specifying Source Referent Contents (SRCs) as SI, TI and disparity. Other parameters describe our interventions – Hypothetical Reference Conditions (HRCs) as PSNR, SSIM or VQM coefficients. All the available sequence parameters (potential regressors) are summed up in Tab. 5. The column titled "Depth description" contains four parameters related to content depth. The first three are the quantiles $(d05, d50, d95)$ of disparity distribution. The fourth parameter is disparity dynamic range, defined as $d95 - d05$. The disparity is calculated for a sufficient amount of significant corresponding pixels by the Speeded-Up Robust Features (SURF) algorithm [35]. The last column contains seven coefficients whose linear combination forms the VQM value.

| 2D parameters | Depth description | 2D metrics | VQM coefficients |
|---|---|---|---|
| SI | d05 | PSNR | si_loss |
| TI | d50 | SSIM | hv_loss |
| | d95 | VQM | hv_gain |
| | dDR | | color1 |
| | | | si_gain |
| | | | contati |
| | | | color2 |

**Tab. 5.** Available objective parameters of the sequences.

We have only 64 samples of MOS, which is the response variable. To avoid over-parametrization, it is necessary to reduce the number of regressors. A good model needs about a hundred observations to one regressor. According to [36], to detect reasonable size effects with reasonable power, 10–20 samples per parameter are needed. The disproportion between the number of potential model parameters and "training" data is also the main reason of that why we focused on linear modeling.

## 5.1 Model Estimation Methods

The simplest and very common model estimation method for the General Linear Model (GLM) is Ordinary Least Squares (OLS). The OLS method minimizes the sum of squared residuals, which are the differences between the observed values and the estimated values of the quantity of interest. In our case, these values are the median of subjectively estimated quality (MOS) and the modeled MOS value. As we cannot exclude the correlation of regressors, Generalized Least Squares (GLS) has been utilized as the model estimation method [36], [37]. There are two criteria on which regressors have been selected into our models: Akaike information criterion and coefficient of determination [38].

The Akaike Information Criterion (AIC) is a measure of the relative quality of statistical models for a given set of data. AIC is based on minimizing the relative information lost when a given model is used to represent the process that generated the data. It sets the proportion between the goodness of fit of the model and the complexity of the model. This level of parsimony is a function of input data sample relevance in a population. The AIC coefficient does not keep any absolute information about model quality, but the model with the lowest AIC is relatively best from the tested set. The coefficient of determination $R^2$ is the proportion of variance explained by the model to the variance of explained (modeled) variable. In the case of linear regression with statistically independent regressors, $R^2$ is the square of the coefficient of multiple correlations between model output and independent (explanatory) variables. The coefficient of determination is increasing with the number of regressors, even if they do not bring other new information. To choose a model with the optimal number of parameters, the adjusted $R^2$ is used. The adjusted $R^2$ ($\approx \widetilde{R^2}$) is the best estimate of the degree of relationship in the basic population. The coefficient $\widetilde{R^2}$ determines how our linear model would describe the population if we had ideal data samples.

The flowchart in Fig. 8 shows the process of setting models (left column), their verification (middle one) and validation (right column). First, the regressors are chosen from Tab. 5 at the base of the criteria mentioned in previous paragraphs. Secondly, the model is set by the GLS model estimation method. The standard deviation per sample ($\sigma$), sometimes in literature called as Root Mean Square Error (RMSE), is calculated. Here, RMSE is the ideal point estimation of $\sigma$.

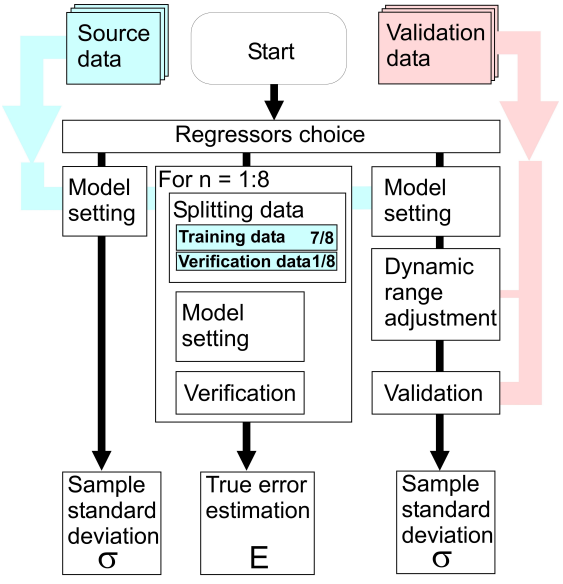



**Fig. 8.** Model setting / verification and validation process.

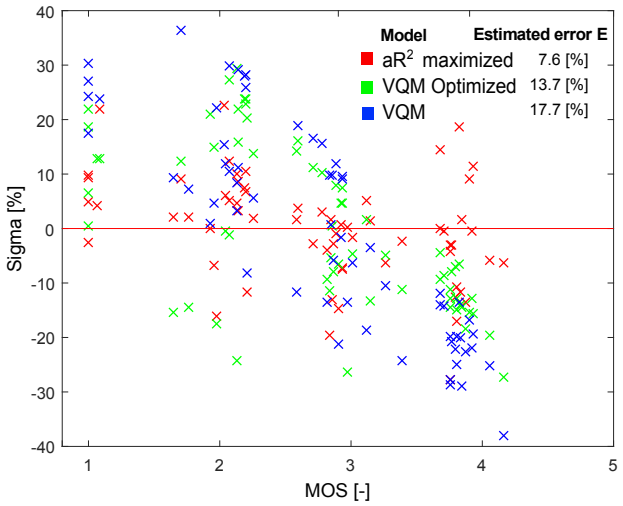



**Fig. 9.** Scatter diagram - residual plot of relative error of MOS.

In the case of model verification, the dataset is randomly divided into 8 parts (literature recommends 5–10, a divisor of 64 was chosen). The model is set to training data (7/8 subpart of original data) and $\sigma_1$ is calculated from verification data (1/8 subpart). After 8 repetitions, the arithmetic mean value of $\sigma_{1-8}$ is calculated, called true error estimation ($E$). Figure 9 shows a residual plot, the scatter plot of verification samples deviations. It demonstrates how the observed values differ from the point of best fit. We can obtain a good overview about model bias and homoscedasticity.

## 5.2 Validation

The right column of the flowchart in Fig. 8 describes the process of model validation. For this purpose, a validation dataset has been added - other video sequences than those used in the subjective test. The validation videos (SRC 1–8) come from RMIT3DV - an uncompressed stereoscopic 3D HD video library. This database has been provided by RMIT University in Melbourne [21]. From the sequences, those whose (potential regressors) parameter values are within the range of the original data values have been chosen. As validation data (SRC 1–8), the sequences 3D_01, 3D_03, 3D_05, 3D_16, 3D_17, 3D_29, 3D_42, 3D_48 were used. The HRC applied on selected sequences was HEVC with four levels of compression ratio (2x [250, 500, 750, 1000] kbps). The validation data is fully independent. The subjective tests have been done with other respondents. Once again, the ACR subjective method was used. Furthermore, the same display technology and test environment have been used.

The dynamic range adjustment is the second step done with the set of model. The full-reference objective video quality metrics as SSIM, VQM, Moving Pictures Quality Metric (MPQM) [39], Noise Quality Measure (NQM) [39] tend to be global QoE models. The generality of the metrics goes against accuracy, even in very complex models. Our goal was to make the most accurate model with limiting data amounts.

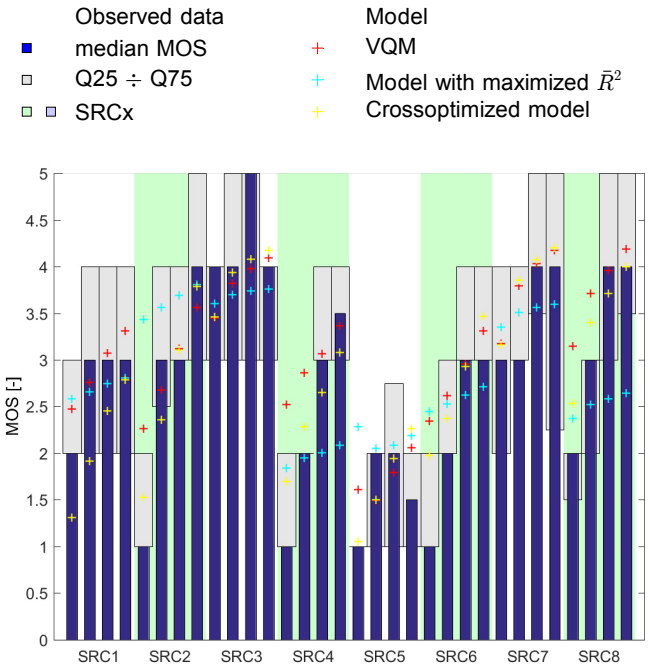



**Fig. 10.** Validation chart - The MOS values of validation data and their model's predictions.

Although the respondents are trained to set their quality dynamic range, they tend to utilize the full range of the QoE scale. This is the reason, why we decided to adjust the dynamic range of our model to validate the data optimally [24, 40, 41].

Validation is done by calculating the standard deviation of the model results and MOS values. The bar graph in Fig. 10 shows MOS values and a gray box containing 50 % of voted quality values. Three various point estimations of MOS, as the three corresponding linear models results, are plotted as color cross marks. The colored background refers to the validation content SRCs 1–8. Each colored surface contains the MOS values of four HRCs applied on one sequence.

## 5.3 Created Models

The details of the models are described in this subsection. Each block of text describes an individual model including its properties and differences from others. Table 6 sums up the model's accuracy and verifications. Correlation coefficient was calculated from original and validation data, therefore, they did not correspond to results from Tab. 4. The $\sigma$ denotes the standard deviation per sample. It is calculated for the original (training) data, verification data and through both datasets (designated $\sum$). The standard deviation has not been calculated for the PSNR metric. This metric does not have defined range. Unlike, for example SSIM metric, where the value ranges from 0 to 1. The standard deviation for model VI. cannot be calculated separately for original and validation data because both datasets are used as the input data of the model.

| MODEL | Description | Regressors | $\widetilde{R^2}$ [-] | $\sigma$ / sample [%] | | | PCC [-] |
|---|---|---|---|---|---|---|---|
| | | | | Original data | Validation data | $\sum$ | |
| PSNR | Generic PSNR | PSNR | 0.376 | - | - | - | 0.411 |
| SSIM | Generic SSIM | SSIM | 0.489 | 46.54 | 43.09 | 45.4 | 0.547 |
| VQM | Generic VQM | VQM | 0.592 | 32.06 | 15.65 | 26.6 | −0.567 |
| I. | VQM optimized | si_loss, hv_loss, hv_gain, color1, si_gain, contati, color2 | 0.577 | 13.8 | 25.51 | 17.7 | 0.631 |
| II. | AIC minimized | PSNR, SI, TI, color2, d50 | 0.842 | 8.57 | 26.37 | 14.5 | 0.713 |
| III. | $aR^2$ maximized | VQM, TI, si_gain, d95 | 0.847 | 8.51 | 21.00 | 12.7 | 0.811 |
| IV. | VQM adjusted | VQM | 0.249 | 19.36 | 13.94 | 17.6 | 0.671 |
| V. | Crossoptimized | TI, hv_gain, color1 | 0.276 | 19.65 | 12.71 | 17.3 | 0.675 |
| VI. | Full-data model | SI, si_loss, si_gain, contati, color2, d05, d50, dDR | 0.772 | - | | 10.9 | 0.889 |
| VII. | Original-data model | SI, si_loss, si_gain, contati, color2, d05, d50, dDR | 0.682 | 12.13 | 22.36 | 18.1 | 0.695 |

**Tab. 6.** Models description, their standard deviation per sample and Pearson Correlation Coefficient.

***Model VQM*** – is a classic VQM, according to the recommendation. It serves us as the reference for other models, due to the fact that it had the highest precision from the general metrics. The deviation per sample is more than two times higher for our original data than for the validation dataset. This indicates that our original dataset is very heterogeneous, which it really is (original sequences are comprised of two different databases).

***Model I*** – is a linear combination of VQM coefficients, optimized for the original data. There are two aspects to demonstrate this model. First, VQM could be improved by training on particular data. The original coefficients of the VQM metric have been established for general 2D video sequences and their analog/digital distortions caused by transmission/broadcasting. There was improve of the VQM model by training it on a specific type of data (stereoscopic 3D sequences in HD resolution). Secondly, there was a lack of data samples to do this VQM optimization properly. The model is overparameterized and loses its generality, which is manifested by the increase of deviation on the validation dataset.

***Model II*** – it has been estimated for our original data. The regressors have been chosen for minimum AIC of the model. More precisely, the AICc coefficient has been minimized, which is preferable in the case of small amounts of data samples (less than 40 samples per 1 coefficient of the model). The coefficients of a linear model for the original data are $[7.72, 3.47, 2.46, 1129, -32.10]$. They are ranked in the same order as the regressors in the Tab. 6. This model is quite well-fitted to our data which is indicated by the adjusted coefficient of determination. However, the validation of this model shows an estimated error value 26%, which is more than in the case of the optimized VQM model. Validation with an independent dataset confirms the concerns that this model is not general enough.

***Model III*** – it includes the regressors which have been chosen based on criteria of the maximum adjusted coefficient of determination. Validation of this model brings better results than the previous one. A scatter plot also indicates good homoscedasticity, even if one of the model coefficients is VQM. The coefficients of regressors for the original data are $[-99.162, 0.844, 470.542, 35.254]$. The result of validation of this model is not cogent (the sample standard deviation is 21%). The author's opinion is that it may be due to the difference between the datasets.

***Model IV*** – it is the VQM value, whose dynamic range is adjusted separately to both datasets. The same process of dynamic range adjustment is done for validation data of models I.–V. The dynamic range of the modeled data (MOS values) is the additive information of the model. So, the model's results should be compared with adjusted MOS values to get relevant information about the model's selective accuracy.

***Model V*** – it has been established to improve the validation dataset results. On the other hand, is not desirable to lost the benefit of two independent datasets which provide information about the generality of the model. The cross-optimization process has been done. From all the models (all the regressor combinations) this have been set to the original data, the algorithm chose the one which has the lowest sample standard deviation for validation data. The algorithm of cross optimization chose the regressors, which have the most similar influence on the modeled MOS value over both datasets. The resulting model is not optimal for any of the datasets, nor for the conjugate dataset, but we do not lose the possibility of validation.

***Model VI*** – it is the optimal model for the conjugate dataset (which includes both original and validation data). The model has been set to a minimum of sample standard deviation and 8 regressors. Although 8 regressors is a reasonable large value for 96 sample of the conjugate dataset, we can deduce nothing about the generality of this model.

***Model VII*** – it is the similar model as the "Model VI" but trained only for the original dataset. The model has the same regressors as before mentioned model. It can be observed that the model has very good results for original data, but worse results are for validation data.

# 6. Conclusion

Regarding the above-mentioned facts and obtained objective and subjective scores, answers to the questions from the Sec. 2 are as follows:

1) After a thorough comparison of all objective and subjective scores, it can be concluded that the VQM objective metric best reflects the user's QoE for stereoscopic 3D video content. However, even this metric does not reach a very high correlation with our subjective test. For more details see the results in Tab. 4 which shows a statistical comparison of objective metrics and subjective tests. From the point of view of outlier rate, it can be concluded that our assumption, that the variance of results is larger at the beginning and at the end of the testing session, due to disorientation and fatigue, was wrong. It was proved that in the subjective test, after a significantly better beginning part (first 8 video sequences), the rest of the evaluation has uniform outlier rate. Dependence of the number of outliers on the sequence was not significant. Results of objective metrics and subjective tests are available at https://www.vutbr.cz/www_base/vutdisk.php?i=145778aa4d.

2) For better modeling of our results, seven new models of objective metrics were created. These models have been validated and verified on other stereoscopic 3D video sequences (see Sec 5.2) and compared to the general VQM model. In general, we can state that the models that had the smallest error for our sequences were less accurate on other databases. On the other hand, models that were less accurate had a wider usable scope on other databases. Table 6 lists the most important data of our models, such as model descriptions, regressors, and their standard deviations. *"Model III"* has the highest correlation with MOS for our dataset. This model, compared to a general VQM metric that had 32% deviation, had only 8% standard deviation. When we consider our source data and validation data, then *"Model VI"* had the smallest standard deviation. On the other hand, the most regressors are included in this model and the model is trained on the original input video sequences as well as on validation sequences. The model, which is the most balanced in all areas, is *"Model V*. The standard deviation for the original sequence is one third lower than for the general VQM. The deviation for the validation data is also slightly lower than for the classic VQM. Another benefit is that only three regressors enter the model calculation. *"Model V"*, for these reasons, can be determined as the most appropriate model due to its great versatility and sufficient accuracy.

# Acknowledgments

This work was supported by the MEYS of the Czech Republic No. LD15020 (QOCIES) and by the BUT project No. FEKT-S-17-4426. Research described in this paper was financed by the Czech Ministry of Education within the frame of the National Sustainability Program under grant LO1401. For research, infrastructure of the SIX Center was used.

# References

[1] ZACH, O., SLANINA, M. A Matlab-based tool for video quality evaluation without reference. *Radioengineering*, 2014, vol. 23, no. 1, p. 405–411. ISSN: 1210-2512

[2] WANG, A., et al. QoE-oriented resource allocation for DASH-based video transmission over LTE systems. In *Proceedings of IEEE Vehicular Technology Conference (VTC Spring)*. Sydney (Australia), 2017, p. 1–5. DOI: 10.1109/VTCSpring.2017.8108535

[3] KALLER, O. *Advanced Methods for 3D Video Capturing and Evaluation*. Dissertation thesis, Brno, 2018

[4] ASSUNCAO, P. A., et al. *3D visual Content Creation, Coding and Delivery*. Cham: Springer International Publishing, 2019. Signals and Communication Technology. ISBN: 978-3-319-77842-6. DOI: 10.1109/ICIP.2007.4379956

[5] ITU-T Rec. H.264. *Advanced Video Coding for Generic Audiovisual Services*. ITU, Geneva (Switzerland), 2014.

[6] Fraunhofer HHI. *High Efficiency Video Coding (HEVC)*. [Online] Cited 2017-03-04. Available from: https://hevc.hhi.fraunhofer.de

[7] Fraunhofer HHI. *Multiview High Efficiency Video Coding (MV-HEVC)*. [Online] Cited 2017-03-04. Available from: https://hevc.hhi.fraunhofer.de/mvhevc

[8] POLAK, L., et al. Study of advanced compression tools for stereoscopic video by objective metrics. In *Proceedings of the 26th International Conference on Radioelektronika*. Kosice (Slovak Republic), 2016, p. 268–272. DOI: 10.1109/RADIOELEK.2016.7477357

[9] HANNUKSELA, M. M., et al. Overview of the multiview high efficiency video coding (MV-HEVC) standard. In *Proceedings of the IEEE International Conference on Image Processing (ICIP)*. Quebec City (Canada), 2015, p. 2154–2158. DOI: 10.1109/ICIP.2015.7351182

[10] BALOTA, G., et al. Overview and quality analysis in 3D-HEVC emergent video coding standard. In *Proceedings of the 5th International Conference on LASCAS*. Santiago (Chile), 2014, p. 1–4. DOI: 10.1109/LASCAS.2014.6820260

[11] STANKOWSKI, J., et al. 3D-HEVC extension for circular camera arrangements. In *3DTV-Conference: The True Vision - Capture, Transmission and Display of 3D Video*. Lisbon (Spain), 2015, p. 1–4. DOI: 10.1109/3DTV.2015.7169371

[12] WANG. K., et al. Stereoscopic 3D video coding quality evaluation with 2D objective metrics. *SPIE Stereoscopic Displays and Applications XXIV*, 2013, vol. 8648, p. 86481L–86481L-7. DOI: 10.1117/12.2003664

[13] YASAKETHU, S.L.P., et al. Quality analysis for 3D video using 2D video quality models. *IEEE Transaction on Consumer Electronics*, 2008, vol. 54, no. 4, p. 1969–1976. ISSN: 0098-3063. DOI: 10.1109/TCE.2008.4711260

[14] BOSC, E., et al. Reliability of 2D quality assessment methods for synthesized views evaluation in stereoscopic viewing conditions. In *3DTV-Conference: The True Vision - Capture, Transmission and Display of 3D Video*. Zurich (Switzerland), 2012, p. 1–4. DOI: 10.1109/3DTV.2012.6365457

[15] JOVELURO, P., et al.Perceptual video quality metric for 3D video quality assessment. In *3DTV-Conference: The True Vision - Capture, Transmission and Display of 3D Video*. Tampere (Finland), 2010, p. 1–4. DOI: 10.1109/3DTV.2010.5506331

[16] HAN, Y., YUAN, Z., MUNTEAN, G. M. Extended no reference objective quality metric for stereoscopic 3D video. In *Proceedings of the IEEE International Conference on Communication Workshop (ICCW)*. London (UK), 2015, p. 1729–1734. DOI: 10.1109/ICCW.2015.7247430

[17] LI, J., BARKOWSKI, M., LE CALLET, P. Validation of reliable 3DTV subjective assessment methodology - Establishing a ground truth database. *VQEG eLetter*, 2014, vol. 1, no. 2, p. 33–35.

[18] TAVAKOLI, S., et al. Subjective quality study of adaptive streaming of monoscopic and stereoscopic video. *IEEE Journal on Selected Areas in Communications*, 2014, vol. 32, no. 4, p. 684–692. DOI: 10.1109/JSAC.2014.140402

[19] PALANIAPPAN, R., JAYANT, N., MANE, P. Visual quality in stereoscopic 3DTV. In *Proceedings of the Conference on Signals, Systems and Computers (ASILOMAR)*. Pacific Grove (USA), 2012, p. 726–728. DOI: 10.1109/ACSSC.2012.6489107

[20] DOMANSKI, M., et al. Poznan multiview video test sequences and camera parameters. *ISO/IEC JTC1/SC29/WG11 MPEG 2009/M17050*, 2009.

[21] CHENG, E., et al. RMIT3DV: Pre-announcement of a creative commons uncompressed HD 3D video database. In *Proceedings of the International Workshop QoMEX*. Yarra Valley (Australia), 2012, p. 212–217. DOI: 10.1109/QoMEX.2012.6263873

[22] ITU-T Rec. P.910. *Subjective Video Quality Assessment Methods for Multimedia Applications*. ITU, Geneva, (Switzerland), 2008.

[23] WANG, Z., BOVIK, A. C., SHEIKH, H. R., SIMONCELLI, E. P. Image quality assessment: From error visibility to structural similarity. *IEEE Transaction on Image Processing*, 2004, vol. 13, no. 4, p. 600–612. ISSN: 1057-7149. DOI: 10.1109/TIP.2003.819861

[24] PINSON, M. H., WOLF, S., A new standardized method for objectively measuring video quality. *IEEE Transaction on Broadcasting*, 2004, vol. 50, no. 3, p. 312–322. DOI: 10.1109/TBC.2004.834028

[25] ITU-R BT.500-13. *Methodology for the Subjective Assessment of the Quality of Television Pictures*. ITU, Geneva, (Switzerland), 2012.

[26] ITU-R BT.2022. *General Viewing Conditions for Subjective Assessment of Quality of SDTV and HDTV Television Pictures on Flat Panel Displays*. ITU, Geneva, (Switzerland), 2012.

[27] SLANINA, M., et al. Testing QoE in different 3D HDTV technologies. *Radioengineering*, 2012, vol. 21, no. 1, p. 445–454.

[28] WOOLSON, R. F. Wilcoxon signed-rank test. *Wiley Encyclopedia of Clinical Trials*, 2007. DOI: 10.1002/9780471462422.eoct979

[29] KRASULA, L., et al. On the accuracy of objective image and video quality models: New methodology for performance evaluation. In *Proceedings of the International Conference on Quality of Multimedia Experience (QoMEX)*. Lisbon (Spain), 2016, p. 1–6. DOI: 10.1109/QoMEX.2016.7498936

[30] NARWARIA, M., et al. Data analysis in multimedia quality assessment: Revisiting the statistical tests. *IEEE Transactions on Multimedia*, 2018, vol. 20, no. 8, p. 2063–2072. DOI: 10.1109/TMM.2018.2794266

[31] MATLAB Documentation. [Online] Cited 2017-03-04. Available from: http://www.mathworks.com/help/matlab/

[32] LEMESHKO, S. B., Distribution of statistics of Chi-square goodness-of-fit tests for small samples. In *Proceedings of the International Conference on Actual Problems of Electronic Instrument Engineering*. Novosibirsk (Russia), 2006, p. 287–287. DOI: 10.1109/APEIE.2006.4292559

[33] ITU-T Rec. P.1401. *Methods, Metrics and Procedures for Statistical Evaluation, Qualification and Comparison of Objective Quality Prediction Models*. ITU, 2012.

[34] OBERTI, F. , et al. ROC curves for performance evaluation of video sequences processing systems for surveillance applications. In *Proceedings of the International Conference on Image Processing*. Kobe (Japan), 1999, p. 949–953. DOI: 0.1109/ICIP.1999.823038

[35] BAY, H., TUYTELAARS, T., VAN COOL, L. SURF: Speeded up robust features. In *Proceedings of the European Conference on Computer Vision*. Graz (Austria), 2006, p. 404–417. DOI: 10.1007/11744023_32

[36] HOLCIK, J., KOMENDA, M., at al. *Mathematics Biology: E-learning Textbook (Matematicka Biologie: E-learningova Ucebnice)*. 1. ed. Brno: Masarykova univerzita, 2015. ISBN 978-80-210-8095-9

[37] HARRELL, F., E. *Regression Modeling Strategies with Applications to Linear Models, in Logistic and Ordinal Regression, and Survival Analysis*. Nashville: Springer, 2001. 571 pages. ISBN 978-3-319-19424-0 DOI: 10.1007/978-3-319-19425-7

[38] AKAIKE, H. A new look at the statistical model identification. *IEEE Transactions on Automatic Control*, 1974, vol. 19, no. 6, p. 716–723. DOI: 10.1109/TAC.1974.1100705.

[39] WANG, Y. Survey of objective video quality measurements. 2006. [Online], Available at: http://digitalcommons.wpi.edu/computerscience-pubs/42

[40] MILOVANOVIC, D., MILICEVIC, Z., BOJKOVIC, Z. MPEG video deployment in digital television: HEVC vs. AVC codec performance study. In *Proceedings of the International Conference on Telecommunications in Modern Satellite, Cable and Broadcasting Services (TELSIKS)*. Nis (France), 2013, p.101–104. DOI: 10.1109/TELSKS.2013.6704900

[41] SULLIVAN, G. J., et al. Overview of the high efficiency video coding (HEVC) standard. *IEEE Transactions on Circuits and Systems for Video Technology*, 2012, vol. 22, no. 12, p. 1649–1668. DOI: 10.1109/TCSVT.2012.2221191

# About the Authors . . .

**Jan KUFA** (1990) received his M.Sc. and Ph.D. degree in Electrical Engineering from Brno University of Technology (BUT) in 2014 and 2018 respectively. His research interests include digital television systems, video image quality, satellite television. Currently he is with at the Department of Radio Electronic, BUT.

**Ondrej KALLER** (1986) received his M.Sc. and Ph.D. degree in Electrical Engineering from Brno University of Technology in 2010 and 2018 respectively. His field of interest includes digital television broadcasting systems. He is focused on 3D video capturing, transmission and interpretation.

**Ondrej ZACH** (1988) received his M.Sc. degree in Electrical Engineering from Brno University of Technology in 2013. At present he is a Ph.D. student at the Department of Radio Electronics, Brno University of Technology. His field of interest is video technology and video coding.

**Ladislav POLAK** (1984) received his M.Sc. degree in 2009, Ph.D. degree in 2013 and Assoc. Prof. in 2018. All in Electronics and Communication from Brno University of Technology, Czech Republic. Currently he is with at the Department of Radio Electronic, BUT. His research interests are Digital Video Broadcasting standards, wireless communication systems, signal processing, video image quality evaluation and design of subjective video quality methodologies. He has been an IEEE member since 2010.

**Tomas KRATOCHVIL** (1976) received his M.Sc. degree in 1999, Ph.D. degree in 2006, Assoc. Prof. in 2009 and Full. Prof. in 2016, all in the Electronics and Communications program from Brno University of Technology. He is currently the Head of the Department of Radio Electronics, Brno University of Technology. His research interests include digital television and audio broadcasting, its standardization and video and multimedia transmission including video image quality evaluation. He has been an IEEE member since 2001 and IEEE senior member since 2016.