Content characterization for bitrate estimation in live video compression and delivery

Shaymaa Al-Juboori and Maria G. Martini
Wireless and Multimedia Networking Research Group,
School of Computer Science and Mathematics,
Kingston University, London, UK
m.martini@kingston.ac.uk

***Abstract*—Compressing video sequences characterized by different content complexity results in different compression bitrates for the same quality level or in different quality levels for the same bitrate; for instance it is well known that content with high spatial complexity and/or high motion requires high bitrates for compression with adequate quality. To address this, per-title optimization is used recently (e.g., by Netflix) to generate appropriate rate-quality representations for different Video on demand (VoD) content to be streamed via adaptive video streaming. However, this cannot be adopted for live video streaming as it requires encoding (multiple times) each video content. Spatial Information (SI) and Temporal Information (TI) have been often used as an indicator of video complexity, for instance for preparing and describing content for video quality assessment tests, and for rate-distortion modeling. However, it has been questioned recently if different metrics could lead to a better estimation of "compressibility" of video. In this paper we compare multiple metrics in terms of their ability to estimate "compressibility". This supports quality-rate estimation and the possibility to create appropriate "quality ladders" (different quality representations) for adaptive live video streaming.**

***Index Terms*—Spatial Information, Temporal Information, Spatial Complexity, Temporal Complexity, Live Video Streaming**

I. Introduction

The measurement of a scene complexity can be used to determine the expected data rate after compression and hence the bandwidth requirement for diverse content types. In fact, more spatially and temporally complex videos require a higher data rate to achieve a satisfactory quality. Measuring the scene complexity plays an important role in key applications ranging from the design of video quality metrics well representative of the quality experienced by the actual users [1][2] to the clustering and classification of different video sequences [3], and data-rate modelling and adaptation [4][5][6]. The metrics used to measure scene complexity range from subjective complexity measures [7] (not suitable for live operation and optimization) to diverse objective metrics [8]. The Spatial Information (SI) of an image [9], as a measure of edge energy, is one of the most widely-used metrics for scene complexity estimation. SI and Temporal Information (TI), as defined by ITU-T Rec. P.910 [10] as an approximate measure of video content complexity, have been widely used in the field of quality assessment, in particular for the selection of the video content to be used for the subjective tests, that should be representative of different complexity classes. For instance, in [4] scene complexity information is used in terms of the spatial content of frames and temporal information is calculated between consecutive frames to derive a rate-distortion model for video sequences. The authors in [11] measure the video quality objectively by utilizing the spatial content of the sequences. The authors in [2], [12] and [1] proposed machine learning based Quality of Experience (QoE) models, where spatial and temporal information values are used along with other influence factors for quality estimation of gaming videos. The applications of spatial and temporal information are not only limited to traditional video sequences and have found application in other fields such as neuromorphic vision. For example, the authors in [6], [5] proposed several spatial information based models to predict the data rate output by Neuromorphic Vision Sensors (NVS).

In [13] we performed a study on spatial and temporal information for different video sequences where we also highlighted that SI and TI do not only depend on video content as often considered, but factors such as resolution, bit depth, compression have an impact on the values of SI and TI for a specific video content. In [14] the authors tested the performance of SI and TI with different pooling methods as complexity measures for video compression.

In order to assess the capability of the complexity metrics to provide a reliable indication on the complexity to compress a specific video, in this paper we compare multiple metrics, also beyond SI and TI, in terms of their ability to estimate "compressibility" of videos. Understanding the best method to estimate video bit-rate would support the creation of appropriate versions of compressed video (different quality representations) for adaptive live video streaming. The rest of this paper is organised as follows: in Section II we list and propose metrics based on spatial complexity, whiles in Section III we list and propose metrics based on temporal complexity. In Section IV we briefly describe alternative metrics proposed very recently. The evaluation methodology adopted is described in Section V, while results are presented and discussed in Section VI.

II. Measuring spatial complexity

A. Methods based on Gradient approximation of Intensity Images

In order to quantify the spatial information as a scene complexity metric, we can consider the mean, standard deviation and root-mean-square of the gradient magnitude image G

[5]. These metrics are computed across all the pixels of the image. Let M and N be the number of rows and columns respectively. The metrics are mathematically expressed as:

$$SC_{\text{mean}} = \frac{1}{M \cdot N} \sum_{j=1}^{N} \sum_{i=1}^{M} G_{i,j} \tag{1}$$

$$SC_{\text{rms}} = \sqrt{\frac{1}{M \cdot N} \sum_{j=1}^{N} \sum_{i=1}^{M} G_{i,j}^2} \tag{2}$$

and

$$SC_{\text{std}} = \sqrt{\frac{1}{M \cdot N} \sum_{j=1}^{N} \sum_{i=1}^{M} (G_{i,j} - SC_{\text{mean}})^2} \tag{3}$$

where SC_{mean}, SC_{rms} and SC_{std} are the mean, root-mean-square and standard deviation based spatial content metrics representing the edge energy of the intensity image *I*.

The gradient of an image specifies the directional change in the intensity of an image. The main step in computing the gradient approximation of an image is to convolve the intensity image with a small finite filter known as kernel. Three well known methods to calculate such gradient are Sobel [15], Prewitt [16] and Roberts [17] filters.

The Spatial Index (SI) complexity metric [18] uses Sobel filtering. The statistic used for spatial pooling is the standard deviation of the magnitude of spatial information ($SI_{std} = \sqrt{\frac{1}{P}\sum(SI_p - SI_{mean})^2}$, where P is the number of pixels in the image. For video sequences, ITU-T Rec. P.910 [10] defines spatial information as:

$$SI = max_{time}\Big\{SI_{std}\Big\}. \tag{4}$$

According to (4), SI_{std} is computed for each of the frames in the video sequence and the maximum of SI_{std}, among all the frames, is taken (over the whole time duration of the sequence).

III. Measuring temporal complexity

A. Temporal Information (TI)

ITU-T Rec. P.910 [10] defined temporal information as:

$$TI = max_{time}\Big\{std[M_p^n]\Big\} \tag{5}$$

$$M_p^n = F_p^n - F_p^{n-1} \tag{6}$$

where M_p^n is the pixel intensity difference between F_p^n, current frame n, and F_p^{n-1}, previous frame $n-1$. For the difference frame the standard deviation is applied across all the pixels. According to (5), the standard deviation of M_p^n is computed for every frame and the maximum is taken over the entire time duration of the video sequence.

IV. Recently proposed spatial and temporal complexity indexes

The Video Complexity Analyzer (VCA) [19] [20] [21] is an open source video complexity analyzer that predicts Spatial Complexity (SC) and Temporal Complexity (TC) (denoted as E and H, respectively) for each frame, video segment, and video. We note that in the rest of this paper we will use SC and E, as well as TC and H, interchangeably.

To calculate SC and TC the following is first calculated:

$$H_{p,k} = \sum_{i=0}^{w-1} \sum_{j=0}^{w-1} e^{|(\frac{ij}{w^2})^2 - 1|} |DCT(\text{i,j})| \tag{7}$$

where k is the block address in the $p-th$ frame, $w \times w$ pixels is the size of the block, and $DCT(i,j)$ is the $(i,j)-th$ DCT component when $i+j > 1$, and 0 otherwise.

The SC is then defined as follows:

$$E = \sum_{k=0}^{C-1} \frac{H_{p,k}}{C \cdot w^2} \tag{8}$$

Where C represents the number of blocks in frame p.

The temporal complexity feature is defined as the block-wise sum of the sum of absolute differences (SAD) in reference to the texture energy of each frame (p) compared to its previous frame $(p-1)$.

$$H = \sum_{k=0}^{C-1} \frac{SAD(H_{p,k}, H_{p-1,k})}{C} \tag{9}$$

V. Evaluation Methodology

A. Datasets

We consider compression artefacts produced by standardized video codecs (H.264, HEVC/H.265), available as open source.

In order to support reproducibility of the results, we conducted experiments on publicly available datasets. The datasets considered for the studies in this paper are described in the following.

BVI-HD [22] is a high definition video quality database that includes 32 references and 384 distorted video sequences (compressed via HEVC), as well as subjective assessments. Six different QP values were considered for HEVC compression (from 22 to 47 with an interval of 5). The sequences have frames rates of 60 fps, 30fps, 50 fps. The duration of the video sequences was truncated to 5s. We consider in our study the subset of sequences at 60fps. Subjective research was conducted to establish the range of quantisation parameters included in the database. Using a double stimulus test approach, the subjective opinion scores for all 384 distorted videos were obtained from a total of 86 individuals. Using a double stimulus test approach, the subjective opinion scores for all 384 distorted videos were obtained from a total of 86 individuals. Figure 1 reports a sample frame for each of the videos in the dataset.

GamingVideoSET [23] is a dataset that contains 24 uncompressed gaming video sequences with a duration of 30 seconds, a resolution of 1080p, and a frame rate of 30 frames

Fig. 1: Examples frames for the videos in BVI-HD dataset [22].

.

Fig. 2: Examples frames for the videos in GamingVideoSet [23].

.

per second, for researchers working on gaming video quality evaluation. Figure 2 reports a sample frame for each of the videos in the dataset. Furthermore, the data set contains subjective quality evaluation scores for 90 video sequences created by encoding six distinct gaming videos in 15 resolution-bitrate pairings using the H.264/MPEG-AVC coding standard (x264) at three resolution and five bitrates each.l A total of 576 distorted videos in MP4 format, obtained by encoding the videos in 24 different resolution-bitrate pairs, and their objective quality assessment results (average and per-frame) using three video quality assessment metrics, are also included in the dataset, in addition to the reference videos. Since the duration of the sequences is too long for the type of study in this paper, we shortened each sequence to 3 seconds (trying to exclude scene cuts in the selected portion).

B. Content characterization metrics compared

Table I shows the details of the metrics compared in this study. Spatial complexity metrics are listed in the first part, followed by temporal complexity metrics.

C. Evaluation method for content charaterization metrics

We restrict our analysis to short duration video sequences of YUV planar colorspace with 4:2:0 chroma subsampling (YUV420) which is the most widely used chroma subsampling scheme across all video streaming and broadcast applications.

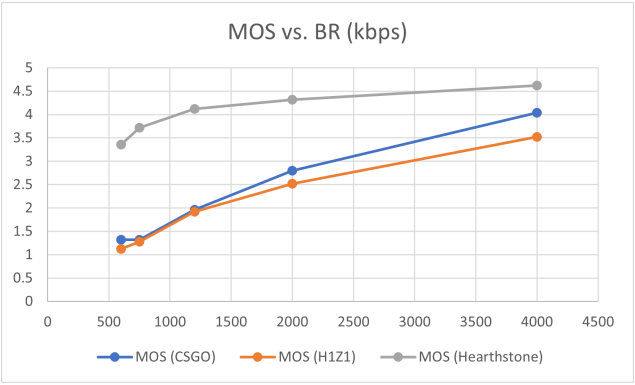


Fig. 3: MOS vs. BR (kbps) for three example contents in [23]
.

TABLE I: Summary of the coding complexity metrics compared

Metric	Definition
SI	$max_{time}\left\{SI_{std}\right\}$
$SC_{meanSobel}$	$mean_{time}\{\frac{1}{M \cdot N}\sum_{j=1}^{N}\sum_{i=1}^{M} G_{i,j\ Sobel}\}$
$SC_{meanPrewitt}$	$mean_{time}\{\frac{1}{M \cdot N}\sum_{j=1}^{N}\sum_{i=1}^{M} G_{i,j\ Prewitt}\}$
$SC_{meanRoberts}$	$mean_{time}\{\frac{1}{M \cdot N}\sum_{j=1}^{N}\sum_{i=1}^{M} G_{i,j\ Roberts}\}$
E [19]	See eqn. (8)
σ^2_{mean}	$mean_{time}\{\sigma^2_Y\}$
σ^2_{max}	$max_{time}\{\sigma^2_Y\}$
σ^2_{min}	$min_{time}\{\sigma^2_Y\}$
σ^2_1	$\sigma^2_{Y_1}$
σ^2_{std}	$std_{time}\{\sigma^2_Y\}$
TI	$max_{time}\left\{std[M^n_p]\right\}$
H [19]	See eqn. (9)

As defined in ITU-T Rec. P.910 [10], SI and TI calculations are performed only on the luminance (Y) channel of the YUV colorspace.

We used MATLAB to read the YUV videos and then performed all SI and TI calculations on the Y channel for all the frames of the YUV videos. In some cases SI and TI values were reported in the datasets. We note that we considered our SI/TI calculation results in the case where we found a discrepancy between our results and those in the dataset.

FFmpeg [24] was used to calculate the Video Multimethod Assessment Fusion (VMAF), Peak Signal to Noise Ratio (PSNR), and Structural Similarity (SSIM) video quality metrics when not provided in the datasets.

D. Performance indicators

In order to select the best metrics to estimate ”compressibility” or ”compression complexity” of a video content we considered compression at different bitrates and the video quality associated to each of them, drawing a quality vs. bitrate curve. Quality was measured based on SSIM, VMAF and/or Mean Opinion Score (MOS). An example of MOS vs. bitrate curves for different contents with different complexity is reported in Figure 3 for content in the dataset [23]. We then calculated for each content the area under the curve (K). Considering that, given a fixed bitrate, lower values of quality are associated to higher complexity of the content in terms of ”compressibility”, we consider the area under the quality-rate curve as an index of ”compression complexity” or ”compressibility”. This is in line with the Bjontegaard delta method used to compare different codecs [25].

In order to be able to compare the area under the curve for different contents, we restricted the curves to a fixed bitrate range. When the data in the dataset was provided for fixed QP values rather than fixed bitrate values, we extrapolated data missing in curves (for instance once a curve saturated to the top VMAF value, we assumed the same top quality value also for the higher bitrates in the range) to cover the selected bitrate range.

We then evaluate the correlation of this measure of ”compressibility” (requiring encoding all the contents at all bitrates) with the simpler indexes listed in Table I above, in terms of Pearson Linear Correlation Coefficient (PLCC), the Kendall Rank Correlation Coefficient (KROCC) and Spearman’s Rank Correlation Coefficient (SROCC).

VI. RESULTS

We report here the results obtained for the BVI-HD dataset, considering VMAF as quality metric.

Since the videos in the dataset were obtained for fixed QP values rather than fixed bitrate values, we extrapolated data missing in curves as mentioned above. We also removed all values corresponding to VMAF lower than 40 in order to align the curves and facilitate the comparison.

Figures 4 and 5 report the correlation values between area under the quality/bitrate curves (for the VMAF video quality metric) and complexity metrics for the BVI-HD dataset. For the second figure, we removed videos with scene cuts, since the purpose of this study is to study the complexity of video contents per scene. Scene cuts would introduce major discontinuities resulting in higher local (in temporal domain) bitrates since motion compensated prediction would fail at the scene cut location. We note that we only compared videos with the same resolution (HD) and same frame rate (60 fps). We observe that, as expected, the complexity metrics usually correlate negatively with the ”compressibility” parameter K calculated as area under the quality-rate curve. Indeed, a higher K means that the video sequence can easily be compressed / reach low bitrates with high quality, hence its complexity for compression is low. The temporal index and temporal complexity in [20] are the best performing metrics for this dataset. Considering spatial complexity only, the spatial complexity metric in [20] and $SC_{meanPrewitt}$ are the best performing metrics for this dataset.

Figure 6 shows the scatter plots complexity metric vs. area under the VMAF vs. rate curve (K_{VMAF}) for two sample spatial complexity metrics ($SC_{mean\ Roberts}$), E); Figure 7 shows the scatter plots complexity metric vs. K_{VMAF} for two complexity metrics related to variance in luminance values, pooled in the time domain via mean and max; finally, Figure 8 shows the scatter plots complexity metric vs. K_{VMAF} for two temporal complexity metrics (TI, H).

Performing tests on different datasets (not reported here for space constraints) we observed that the best performing metrics are not the same in the different datasets as the

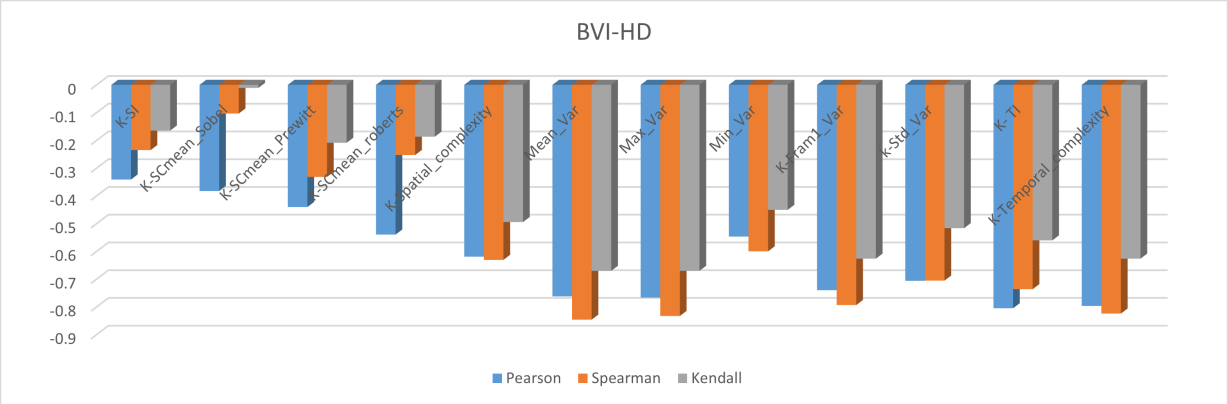


Fig. 4: Bar plot depicting the correlation between complexity measures and K_{VMAF} (area under the VMAF-bitrate curve), BVI-HD dataset.

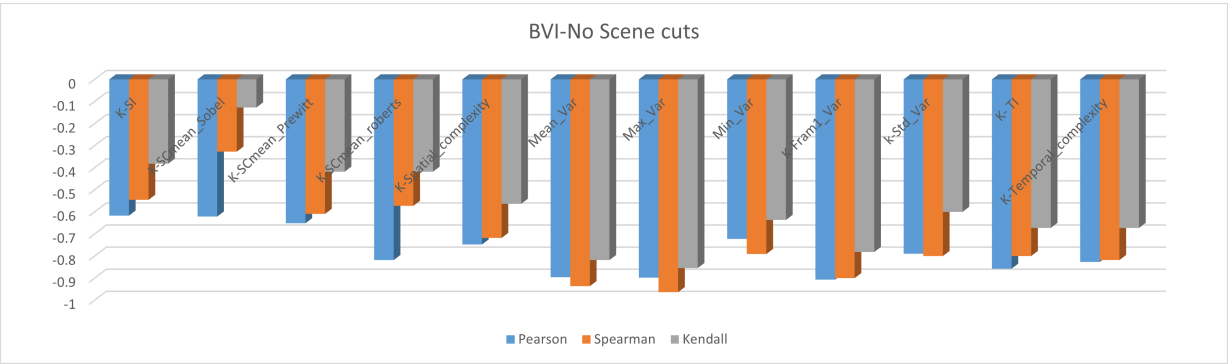


Fig. 5: Bar plot depicting the correlation between complexity measures and K_{VMAF} (area under the VMAF-bitrate curve), BVI-HD dataset no scene cuts.

performance also depends on the type of content / genre (e.g., animation, gaming videos, cinema movies, news, sport).

VII. Conclusion and Future Work

The comparison of multiple metrics in terms of their ability to estimate "compressibility" has shown that alternatives to the popular SI and TI metrics exist. Results with different types of content/datasets can help further generalising the results and in an extended version of this work we will present more detailed results for more contents in different datasets and also including different metrics for content characterization.

References

[1] N. Barman, E. Jammeh, S. A. Ghorashi, and M. G. Martini, "No-Reference Video Quality Estimation Based on Machine Learning for Passive Gaming Video Streaming Applications," *IEEE Access*, vol. 7, pp. 74 511–74 527, June 2019.

[2] S. Zadtootaghaj, N. Barman, S. Schmidt, M. G. Martini, and S. Möller, "NR-GVQM: A No Reference Gaming Video Quality Metric," in *2018 IEEE International Symposium on Multimedia (ISM)*, Taichung, Taiwan, Dec 2018, pp. 131–134.

[3] S. Zadtootaghaj, S. Schmidt, N. Barman, S. Möller, and M. Martini, "A classification of video games based on game characteristics linked to video coding complexity," in *Network and Systems Support for Games (NetGames), 2018 16th Annual Workshop on*, Amsterdam, Netherlands, June 2018.

[4] A. Haseeb, M. G. Martini, S. Cicalo, and V. Tralli, "Rate and distortion modeling for real-time MGS coding and adaptation," in *2012 Wireless Advanced (WiAd)*. IEEE, 2012, pp. 85–89.

[5] N. Khan and M. G. Martini, "Bandwidth modeling of silicon retinas for next generation visual sensor networks," *Sensors*, vol. 19, no. 8, p. 1751, 2019.

[6] N. Khan and M. Martini, "Data rate estimation based on scene complexity for dynamic vision sensors on unmanned vehicles," in *IEEE International Symposium on Personal, Indoor and Mobile Radio Communications (PIMRC)*, Bologna, Italy, September 2018.

[7] V. Chikhman, V. Bondarko, M. Danilova, A. Goluzina, and Y. Shelepin, "Complexity of images: Experimental and computational estimates compared," *Perception*, vol. 41, pp. 631–647, 2012.

[8] R. Cilibrasi and P. M. B. Vitanyi, "Clustering by compression." *IEEE Transactions on Information Theory*, vol. 51, pp. 1523–1545, 2005.

[9] H. Yu and S. Winkler, "Image complexity and spatial information," in *IEEE International Conference on Quality of Multimedia Experience (QoMEX)*, Klagenfurt, Austria, 2013, pp. 12–17.

[10] ITU-T Rec. P.910, *Subjective video quality assessment methods for multimedia applications*, ITU-T Recommendation, April 2008.

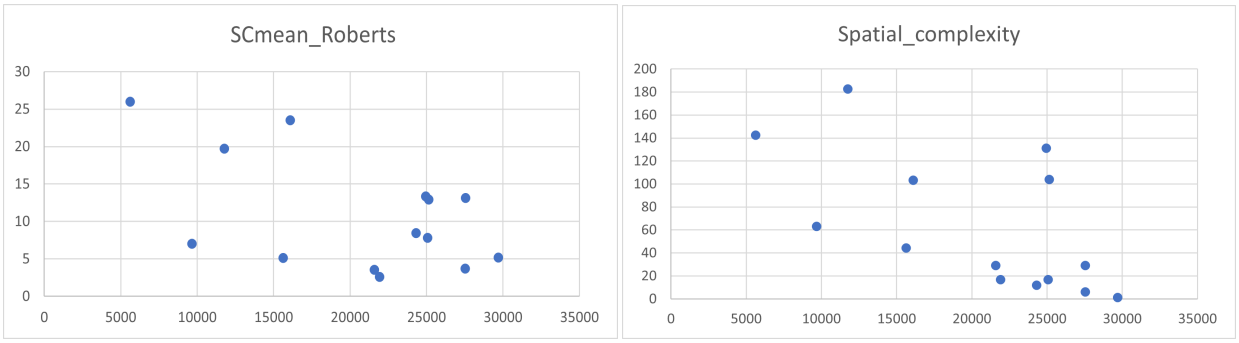
[11] M. H. Pinson and S. Wolf, "A new standardized method for objectively measuring video quality," *IEEE Transactions on Broadcasting*, vol. 50, no. 3, pp. 312–322, Sept 2004.

[12] S. Göring, R. R. R. Rao, and A. Raake, "nofu — A Lightweight No-Reference Pixel Based Video Quality Model for Gaming Content," in *2019 Eleventh International Conference on Quality of Multimedia Experience (QoMEX)*, June 2019, pp. 1–6.

[13] N. Barman, N. Khan, and M. G. Martini, "Analysis of spatial and temporal information variation for 10-bit and 8-bit video sequences," in *2019 IEEE 24th International Workshop on Computer Aided Modeling and Design of Communication Links and Networks (CAMAD)*. IEEE, 2019, pp. 1–6.

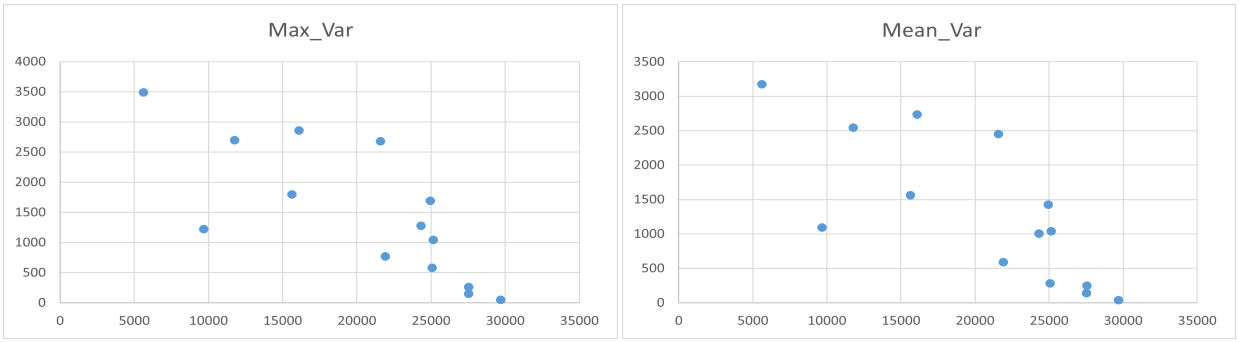
[14] W. Robitza, R. R. R. Rao, S. Göring, and A. Raake, "Impact of spatial and temporal information on video quality and compressibility," in *2021 13th International Conference on Quality of Multimedia Experience (QoMEX)*. IEEE, 2021, pp. 65–68.

[15] I. Sobel and G. Feldman, "A 3x3 isotropic gradient operator for image



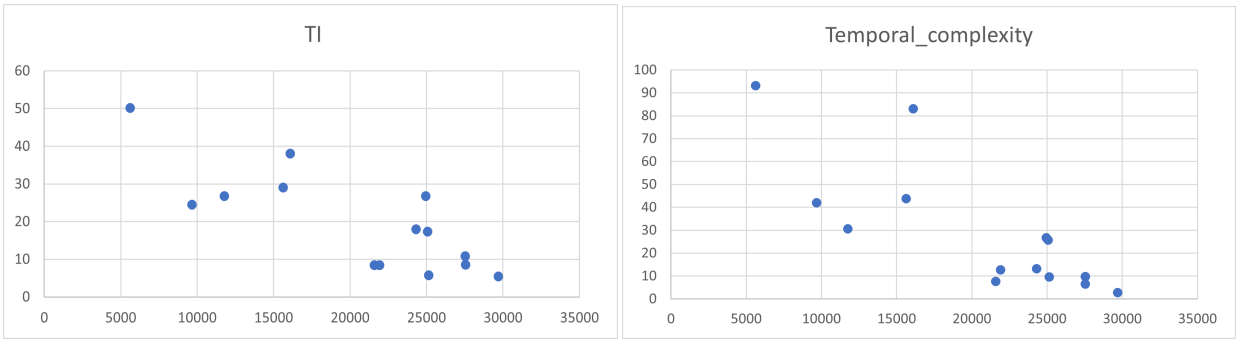
(a) (b)

Fig. 6: Scatter plots for two example spatial complexity measures vs. K_{VMAF} (area under the VMAF-bitrate curve), BVI-HD dataset. (a) $SC_{mean\ Roberts}$; (b) E.



(a) (b)

Fig. 7: Scatter plots for the two best spatial complexity measures vs. K_{VMAF} (area under the VMAF-bitrate curve), BVI-HD dataset. (a) σ^2_{max}; (b) σ^2_{mean}.



(a) (b)

Fig. 8: Scatter plots for two example temporal complexity measures and K_{VMAF} (area under the VMAF-bitrate curve), BVI-HD dataset. (a) TI; (b) H.

processing, presented at a talk at the Stanford Artificial Project," in *Pattern Classification and Scene Analysis*, R. Duda and P. Hart, Eds. New York: John Wiley & Sons, 1968, pp. 271–272.

[16] J. M. S. Prewitt, "Object enhancement and extraction," in *Picture Processing and Psychopictorics*, B. Lipkin and A. Rosenfeld, Eds. New York: Academic Press, 1970, pp. 75–149.

[17] L. G. Roberts, "Machine perception of three-dimensional solids," Ph.D. dissertation, Massachusetts Institute of Technology, 1963.

[18] S. Wolf and M. H. Pinson, "Spatial-temporal distortion metric for in-service quality monitoring of any digital video system," in *Multimedia Systems and Applications II*, vol. 3845. SPIE, 1999, pp. 266–277.

[19] V. V. Menon, C. Feldmann, H. Amirpour, M. Ghanbari, and C. Timmerer, "VCA: Video complexity analyzer," 2022. [Online]. Available: https://github.com/cd-athena/VCA

[20] V. V. Menon, H. Amirpour, M. Ghanbari, and C. Timmerer, "Efficient bitrate ladder construction for live video streaming," in *Proceedings of the 1st Mile-High Video Conference*, 2022, pp. 99–100.

[21] ——, "Opte: Online per-title encoding for live video streaming," in *ICASSP 2022 - 2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2022, pp. 1865–1869.

[22] F. Zhang, F. M. Moss, R. Baddeley, and D. R. Bull, "BVI-HD: A video quality database for HEVC compressed and texture synthesized content," *IEEE Transactions on Multimedia*, vol. 20, no. 10, pp. 2620–2630, 2018.

[23] N. Barman, S. Zadtootaghaj, S. Schmidt, M. G. Martini, and S. Möller, "GamingVideoSET: a dataset for gaming video streaming applications," in *2018 16th Annual Workshop on Network and Systems Support for Games (NetGames)*. IEEE, 2018, pp. 1–6.

[24] "FFmpeg,," https://ffmpeg.org/, [Online: accessed 12-03-2022].
[25] N. Barman, M. G. Martini, and Y. Reznik, "Revisiting Bjontegaard delta bitrate (BD-BR) computation for codec compression efficiency comparison," in *Proceedings of the 1st Mile-High Video Conference*, 2022, pp. 113–114.