This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TMM.2018.2817070, IEEE Transactions on Multimedia

BVI-HD: A Video Quality Database for HEVC Compressed and Texture Synthesised Content

Fan Zhang, *Member, IEEE*, Felix Mercer Moss, *Member, IEEE*, Roland Baddeley and David R. Bull, *Fellow, IEEE*

***Abstract*—This paper introduces a new high definition video quality database, referred to as BVI-HD, which contains 32 reference and 384 distorted video sequences plus subjective scores. The reference material in this database was carefully selected to optimise the coverage range and distribution uniformity of five low level video features, while the included 12 distortions, using both original High Efficiency Video Coding (HEVC) and HEVC with synthesis mode (HEVC-SYNTH), represent state-of-the-art approaches to compression. The range of quantisation parameters included in the database for HEVC compression was determined by a subjective study, the results of which indicate that a wider range of QP values should be used than the current recommendation. The subjective opinion scores for all 384 distorted videos were collected from a total of 86 subjects, using a double stimulus test methodology. Based on these results, we compare the subjective quality between HEVC and synthesised content, and evaluate the performance of nine state-of-the-art, full-reference objective quality metrics. This database has now been made available online, representing a valuable resource to those concerned with compression performance evaluation and objective video quality assessment.**

***Index Terms*—Subjective quality assessment, BVI-HD video quality database, visual perception, HEVC, synthesis-based compression**

I. Introduction

The technologies to support emerging immersive formats including AR and VR via Head Mounted Displays (HMDs) as well as conventional displays with higher frame rates, spatial resolutions and dynamic ranges, are progressing rapidly. Such formats are becoming increasingly popular as they provide the potential to offer new sustainable multimedia services and experiences. There is significant research activity in all aspects of this technology [1–12], from content design to displays, new immersive audio-visual formats and quality assessment.

With particular relevance to this paper, there is also significant activity in the video compression community, much of it linked to the development of new standards for coding of high spatio-temporal resolutions, high dynamic range content and 360 degree immersive formats [13, 14]. There are a number of key challenges associated within emerging immersive formats such as 360 where 'immersion-breaking' artefacts due to compression (delays, peripheral vision distractions, motion blur/aliasing, temporal and spatial resolution) are very important and must be avoided [15, 16]. Such artefacts are not just annoying in the sense that they reduce visual quality, but in more immersive formats, visual inconsistencies, anomalies or artefacts can have a significant negative impact on user immersion though the coupling to other senses such as the auditory and vestibular systems.

Hence there is a high emphasis on compression performance, particularly with six degree of freedom (6DoF) 360 formats where raw bit rates would exceed 140Gbps (8K per eye at 12 bits, 3 colour channels and 120fps). Delivery bit rates must be at a fraction, typically 15-25Mbps (for Ultra-high Definition video streaming) [17], of this implying compression ratios of many 1000s:1. Therefore new coding techniques are needed to deliver content at manageable bit rates while ensuring that the immersive properties of the format are preserved. This work is timely as a new draft call for proposals for future coding beyond HEVC [18] has now been launched so techniques for addressing this problem are urgently required.

Before widespread adoption, all proposed coding algorithms must be performance tested and compared using databases containing diverse samples of video content. Additionally, video sequences with various levels of distortion, alongside corresponding subjective quality scores, are often employed to evaluate the prediction accuracy of objective quality assessment methods. Since the performance of video codecs and quality metrics are often content-dependent, criteria for the selection of original sequences in subjective databases is very important.

Since the first major subjective video quality database, VQEG FR-TV Phase I [19], was developed, the last decade has seen the publication of numerous video databases for different applications. In most cases, a video database for testing compression performance and evaluating quality assessment methods is expected to be composed of diverse content. Previous research in [20] has reported that the diversity of a database can be measured by the coverage range and uniformity of various video features. These two metrics are therefore often used to assess and compare video content of datasets, and also to facilitate reference material selection.

In most video quality databases, distorted versions of the original reference videos are created based on different levels of compression and transmission loss. The former are usually based on the state-of-the-art video compression standards.

Manuscript drafted at September 2017, revised at January 2018.

The authors acknowledge funding from EPSRC (grant EP/J019291/1 and EP/M000885/1), and the support from BBC R&D. We also would like to thank the Consumer Digital Video Library (CDVL) for the permission to use and modify original sequences from the Video Quality Experts Group (VQEG) HDTV Phase I database, and for the distribution of the BVI-HD database.

Fan Zhang, Felix Mercer Moss, Roland Baddeley and David R. Bull are with the Bristol Vision Institute, University of Bristol, Bristol, UK.

E-mail: fan.zhang@bristol.ac.uk, f.mercermoss@bristol.ac.uk, roland.baddeley@bristol.ac.uk and dave.bull@bristol.ac.uk

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TMM.2018.2817070, IEEE Transactions on Multimedia

However, four years after the publication of the most recent standardised codec [13], very few high definition (HD) video databases containing HEVC compressed content together with opinion score metadata are publicly-available.

In addition, more recent approaches [21–26] that use texture analysis and synthesis to compress video content have demonstrated significant potential for rate-quality performance improvement. This type of codec produces distinct distortions and artefacts that are conspicuously absent from databases currently available.

In this context, this paper presents a novel video database for evaluating video compression algorithms and quality assessment methods. The BVI-HD video quality database[1] contains 32 HD, progressive-scanned reference sequences, selected to maximise video feature coverage and uniformity. Each reference is accompanied by 12 distorted variants: half of these using HEVC and half using a synthesis-based video codec, HEVC-SYNTH.

We conducted subjective experiments to address three specific research topics in video compression and quality assessment.

1) To determine the optimal test range of quantisation parameters in HEVC compression for video quality assessment (VQA) studies.
2) To understand whether there is difference in terms of subjective quality between HEVC compressed and HEVC+synthesis content.
3) To evaluate the performance of objective video quality metrics on the BVI-HD video quality database.

The remainder of this paper is organised as follows. Section II summarises the most recent research in video database development. Section III presents the reference and distorted sequences of the database, while the conducted subjective experiments are described in Section IV. The subjective results of the database are reported and discussed in Section V. Finally, Section VI concludes the paper, and summarises future work.

II. Background

This section discusses the primary considerations for developing video databases, such as source content, sequence length, and test methodology, and provides a brief review on existing video quality databases.

A. Video Content

Irrespective of what a video database is developed for, video content is always important. However, the content selection process, for many existing databases, is based on subjective preference and video availability. This is due to the lack of a standardised approach to video parameterisation and characterisation. In most cases, it is essential to consider the diversity of content as this determines how representative the database is, and consequently, the validity of the evaluation process.

[1]All video sequences and subjective scores of the BVI-HD database is publicly available for downloading at *https://vilab.blogs.ilrt.org/?p=1946*.

Measuring how representative a set of videos is, of a much larger population, is not a trivial task. One popular approach is to characterise the videos in a database using low-level features, based on spatial and temporal information, and to calculate the range and uniformity of feature distributions [20]. While this can be an effective way of ranking database coverage, it does so independently of, and without reference to, contemporary content. After conducting a large-scale analysis of recently-broadcast content [27], we previously identified the distributions of five uncorrelated factors that explain the directions of highest variance in the videos they analysed [28]. Analysing the shape of their distributions provides a convenient approach for comparing the content of a limited number of sequences in a database to the wider, near-infinite population of modern broadcast content.

B. Sequence Length

In a subjective VQA experiment, the duration of each video sequence is important from both practical and theoretical perspectives. It determines the overall experiment time consumed for testing – using shorter video durations offers the opportunity to collect similar volumes of subjective data within a shorter amount of time. The presentation time for moving pictures is recommended by the ITU as ten seconds [29]. However, recent studies in [30, 31] explicitly recommend the use of five second instead of ten in video quality assessment for both single and double stimulus methodologies. This reflects the the average shot length in modern films [32], and shorter clips also produce more critical [33] and consistent [34, 35] test results.

C. Test Methodology

Test strategies using different stimulus presentation methods and rating scales are well documented in [29, 36]. Double-stimulus (DS) and single-stimulus (SS) are two commonly used test methodologies. In a DS experiment, before requesting observers to vote, a pair of sequences are presented, and the difference opinion scores between these two sequences are recorded. If the SS approach is used, all videos are randomly presented to subjects. After viewing a single clip, they are required to provide an absolute quality score. This methodology is more suitable for cases without explicit reference sequences and significantly saves testing time. In comparison to the DS method, SS is more efficient but could suffer context effects. Detailed reviews for various test methodologies can be found in [17, 30].

D. Existing Video Quality Databases

Subjective video databases, designed for the validation of objective quality metrics, in most cases, contain distorted content with coding artefacts and/or transmission errors. It is noted that, since the publication of the latest compression standard (HEVC) [13], few publicly-available HD video databases have included subjective data on HEVC compressed content.

VQEG FR-TV Phase I [19], and LIVE [37] are two commonly used subjective video databases, which contain

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TMM.2018.2817070, IEEE Transactions on Multimedia

sequences at standard definition (SD) resolutions, using H.263 (VQEG), MPEG-2 (VQEG and LIVE), and H.264/AVC (LIVE) codecs to generate distorted content. Other notable contributions include VQEG-HD [38, 39], IRCCyN/IVC 1080i [40], IVP [41], VQEG HDTV Phase I [38, 39], EPFL [42], MMSP-SVD [43], Poly@NYU [44] and BVI-HFR [6] databases. None of these contains HEVC compressed content.

Current subjective databases that feature HEVC test content include CSIQ [45], BVI Texture [46], SJTU 4K-HEVC [47], BVI-HomoTex [48] and Yonsei 4K UHD database [49]. However, the CSIQ and BVI-HomoTex databases only includes low resolution video sequences (480p and 256p respectively), while the content in the BVI Texture database is limited to homogeneous textures. The SJTU 4K-HEVC and Yonsei 4K UHD databases do include HEVC compressed content at 4K (3840×2160) resolution, but the limited number of source (10 in Yonsei 4K UHD) or distorted (60 in SJTU 4K-HEVC) sequences may lead to unreliable results when used to validate objective quality models.

III. The BVI-HD Video Quality Database

This section presents the procedure used to select the 32 reference sequences in the BVI-HD video quality database, and describes how the 384 distorted videos were generated.

A. Reference Sequences

One hundred and thirty-one original uncompressed sequences from the VQEG HDTV Phase I [38, 39], BVI-HD Texture [46] and BVI-HFR [6] databases were included in the initial video selection pool. Each of these progressive-scanned, high definition (HD, 1920×1080) sequences, was truncated from its original length to just five seconds using the method recommended in [30, 31] to maximise temporal consistency and minimise scene cuts.

Algorithm 1 The procedure for source sequence selection. Here $|\mathbf{D}|$ represents the cardinality of set $\mathbf{D}$.

Input:
Candidate video pool: $\mathbf{V} = \{V_1, V_2, \cdots, V_{131}\}$;
Corresponding video features for candidate videos:
$\{\mathrm{SI}_1, \mathrm{SI}_2, \cdots, \mathrm{SI}_{131}\}$, $\{\mathrm{CF}_1, \mathrm{CF}_2, \cdots, \mathrm{CF}_{131}\}$,
$\{\mathrm{MV}_1, \mathrm{MV}_2, \cdots, \mathrm{MV}_{131}\}$, $\{\mathrm{TP}_1, \mathrm{TP}_2, \cdots, \mathrm{TP}_{131}\}$,
$\{\mathrm{DTP}_1, \mathrm{DTP}_2, \cdots, \mathrm{DTP}_{131}\}$
Empty dataset: $\mathbf{D} = \{\}$;
Output:
Full dataset: $\mathbf{D} = \{V_{N_1}, V_{N_2}, \cdots, V_{N_{32}}\}$

1: Move candidate videos with extreme (maximum and minimum) feature values from $\mathbf{V}$ to $\mathbf{D}$;
2: **while** $|\mathbf{D}| < 32$ **do**
3: Calculate the uniformity of $\mathbf{D} \cup \{V_i\}$ for feature SI, $\forall V_i \in \mathbf{V}$;
4: Move the video with the highest SI uniformity level from $\mathbf{V}$ to $\mathbf{D}$;
5: **if** $|\mathbf{D}| \geq 32$ **then**
6: **return** $\mathbf{D}$;
7: **end if**
8: Repeat 3-7 for features CF, MV, TP and DTP;
9: **end while**
10: **return** $\mathbf{D}$

It is noted that while previous video databases have favoured researcher intuition for scene selection, this strategy can be vulnerable to the inclusion of redundant content and can exhibit non-uniform content distribution, particularly within larger sets of reference sequences. Containing 32 unique references, the BVI-HD video database is larger than most; therefore we developed a selection algorithm to maximise the range and uniformity of content within the final set of reference scenes.

In this method, thirty-two source sequences were chosen from these 131 candidate videos based on the procedure to maximise the range and uniformity [20] of five video features: spatial information (SI), colourfulness (CF), motion vector (MV), texture parameter (TP) and dynamic texture parameter (DTP). The detailed definition of SI, CF and MV can be found in [20], while TP and DTP have been described in [30]. The calculation of DTP was based on normalised motion vectors, and the normalisation method is described in [20].

Algorithm 1 describes the source content selection process. After 32 source sequences were chosen, each selected video was examined by eye to identify any unsuitable content, e.g. with scene cuts or computer generated animation (only natural content is considered). When this was the case, the video in question was excluded, and the selection algorithm was applied again from STEP 2 until the final acceptable 32 source sequences were identified. Sample frames from the final selection of 32 source sequences alongside sequence names and indices are shown in Fig. 1, and their feature distributions are illustrated in Fig. 2.

In order to further evaluate the representativeness of the selected set of sequences to contemporary broadcast content, we projected each into the 5-dimensional factor space described in [28], labelled: *Naturalness*, *Movement*, *Brightness*, *Contrast* and *Saturation*. Fig. 3 plots the cumulative distribution curves of these five factors for the BVI-HD video quality database, superimposed over the same curves calculated from 14075 clips of BBC broadcast content (from the BBC Redux database [27]), as reported in [28]. It is observed that for all five factors, the BVI-HD database follows similar distributions compared with BBC Redux samples.

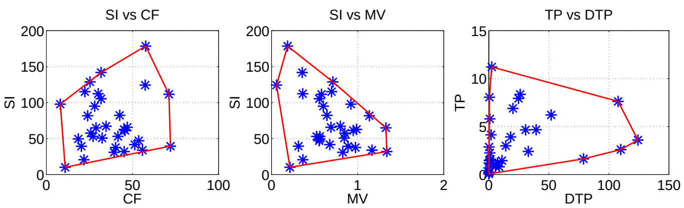
We also applied a two-sample Kolmogorov-Smirnov test (K-S test) [50] to see if there were significant differences between the BVI-HD database and the BBC Redux data [27] for the distributions of all five factors. The K-S statistic values for all five factors are 0.134, 0.183, 0.123, 0.142 and 0.104 respectively (their corresponding p values are 0.589, 0.213, 0.694, 0.509 and 0.864); all lower than the threshold 0.241 for a significant level of 0.05. These results show that, for all five factors, there is no significant difference between the distributions of these two data samples (BVI-HD and BBC Redux) at a 95% confidence interval, which indicate that the chosen video set is representative of contemporary broadcast content.

B. Test Sequence Generation

Each original sequence was distorted by HEVC compression (HM 14.0) using the Random Access configuration. Primary coding parameters include: Main Profile; six different QP values from 22 to 47 with an interval of 5; three successive bi-directional predicted B frames (group of picture size 4);

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TMM.2018.2817070, IEEE Transactions on Multimedia

Fig. 1: Sample video frames from the BVI-HD database source sequences.



(a) SI versus CF (b) SI versus MV (c) TP versus DTP

Fig. 2: Feature distributions of the proposed database.

coding tree unit (CTU) size 64×64; QPOffset values are set to zero.

The QP range used here (22–47) is different from the recommendation in [51, 52], and this was justified by a subjective study on QP range selection described in Section IV-C. It is also noted that the GOP size and QPOffset values employed were different from the HEVC Common Test Conditions in [51, 52]. This is because, by using longer GOP size and various frame level QP values (QPOffset values are non-zero), the visual quality of each frame may change significantly in the temporal domain (this is also content dependent). In order to simplify the case, we have adopted zero QPOffset values and shorter GOP size, which could produce temporally more consistent perceptual quality. The influence of various frame level QP values on overall visual quality will be investigated in our future work.

Due to recent advances in synthesis-based video compression [21–25], synthesised content was also included in this database. The approach described in the appendix was employed to create synthesised results based on the generated HEVC compressed content (for QP 27 and 42 only). In this approach, three quality threshold levels were used to control the synthesis. Therefore, for each reference clip, there are six (3×2) distorted versions with synthesis artefacts.

C. Summary

In total, the final database consisted of 32 reference and 384 (32×12) distorted sequences (each with a duration of five seconds). The distortion coverage of the BVI-HD database is shown in Fig.4 alongside those for three other existing databases, VQEG [19], LIVE [37], and IVP [41], where the most commonly used quality metric, PSNR, is used to predict distortions. It can be observed that the histogram plot for all 384 test sequences of the BVI-HD database offers a wider coverage across the whole PSNR range compared to the three other datasets.

IV. Subjective Experiments

Two subjective studies are reported here: the first was an experiment to determine the appropriate QP range used to generate distorted videos; the second collected opinion scores for each distorted sequence in the proposed database.

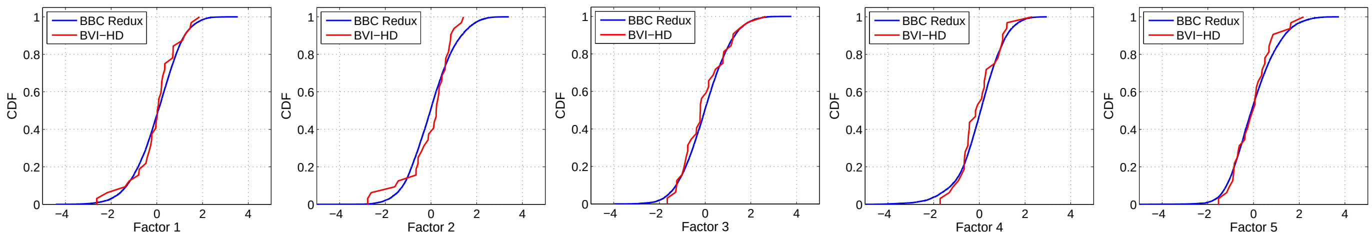
A. Environmental Setup

Both experiments were conducted in a darkened, living room-style environment. The background luminance level was set to 15% of the peak luminance of the monitor used (32 lux) [29]. Video sequences were shown at their native framerates, on a Panasonic BT-4LH310 LCD professional reference monitor, which measures 700×370mm, with a static contrast ratio of up to 1500:1, and with a maximum viewing angle of 178°. The resolutions of the monitor were configured to 1920×1080 (spatially) and 60Hz (temporally). It was connected to a Windows PC running Matlab R2012a and Psychotoolbox 3.0. The viewing distance was set to be three times of the monitor height (1110mm), which is within the recommended range in ITU-R BT.500 [29]. During both experiments, a 9.71″ iPad tablet computer was provided to each participant, and a customised iOS app was used to collect opinion scores.

B. Experimental Procedure

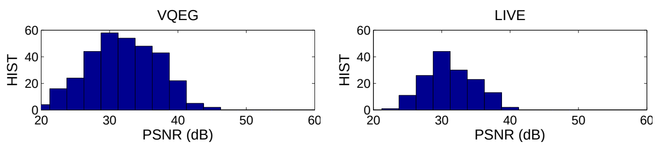
In both experiments, the double stimulus continuous quality scale (DSCQS) [29] methodology was used. In each trial, participants were shown sequence A, then sequence B, followed by a grey screen, at which point participants had unlimited time to respond to two questions issued on the tablet computer

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TMM.2018.2817070, IEEE Transactions on Multimedia



(a) Factor 1 - Naturalness (b) Factor 2 - Movement (c) Factor 3 - Brightness (d) Factor 4 - Contrast (e) Factor 5 - Saturation

Fig. 3: The cumulative distribution function (CDF) curves of five features identified in [28]. Blue curves are calculated from the high density sampling of BBC broadcast content, while red curves are calculated using the reference sequences in the BVI-HD database.



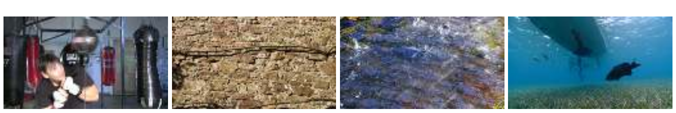
(a) VQEG (b) LIVE (c) IVP (d) BVI

Fig. 4: Distortion coverage and distribution for various databases. The PSNR histogram of test sequences are used here.

which were individually assigned. The first of these asked "*Which video did you perceive as better quality?*". Participants registered their answers by clicking either 'Video A' or 'Video B' on the tablet. The second was phrased "*Please rate the perceived quality of the two videos.*". The sliders next to the visual analogue scales in the iOS app were provided to allow participants to record their answers, and the scales featured evenly-spaced labels reading: *Excellent*, *Good*, *Fair*, *Poor* and *Bad*.

C. Exp. 1: the subjective study on QP range selection

This experiment was conducted to identify a suitable range of QP values for the BVI-HD video quality database. Four HD reference sequences, labelled *Boxing*, *BrickPanning*, *SplashingWater*, and *UnderwaterFish*. *Boxing* and *UnderwaterFish* were chosen from the VQEG HDTV Phase I [38, 39] database, while *SplashingWater* was selected from the BVI Texture [46] database. *BrickPanning* is a newly captured sequence using a professional camera (RED-EPIC). These were all truncated to 5 seconds based on the procedure in [30, 31]. Sample frames are shown in Fig. 5.



Boxing BrickPanning SplashingWater UnderwaterFish

Fig. 5: Sample frames from the source videos in Exp. 1.

Source videos were encoded by an HEVC reference codec (HM 14.0) using six different quantisation parameters (20, 24, 28, 32, 36, 40, 44, 48). Other coding configurations are the same as those given in Section III-B. Six university postgraduate students (three male and three female with an average age of 29) participants, reported with normal or corrected-to-normal vision (visual acuity was confirmed with Snellen charts, and Ishihara chart was used for colour blindness test), viewed and rated each reference and each distorted version three times making a total of 96 (4×8×3) trials following the methodology in Section IV-B.

D. Exp. 2: the BVI-HD video quality database

Data collection for the BVI-HD database was divided into four participant groups: the first group viewed trials containing original, HEVC compressed sequences from the first 16 references (source 1–16); the second group viewed trials containing HEVC compressed sequences from the remaining 16 references (source 17–32); the third group of participants were shown the distorted sequences generated by the synthesis-based codec from source 1–16; and the final group of subjects viewed the synthesised sequences (source 17–32). For each test session, up to three participants simultaneously viewed three blocks of 32-length trials. Each block lasted no more than 15 minutes and participants were given the option of a 5-minute break between blocks. Trials within each block were randomly permutated at the beginning of each session, as were the order of the reference and distorted sequences.

A total of 86 undergraduate and postgraduate students from the University of Bristol with an average age of 26 were financially compensated for their participation in the experiment. After visual acuity and colour blindness were tested using a Snellen chart and a Ishihara chart respectively, participants were given instructions and presented with two training trials containing videos that were not featured in the main experiment. The subjective data in the training session were not collected for subsequent analysis. Twenty-two viewers (10 male and 12 female) were in Group 1; Twenty-one viewers (10 male and 11 female) were in Group 2; Twenty-two viewers (9 male and 13 female) were in Group 3; Twenty-one viewers (11 male and 10 female) were in Group 4.

E. Data Processing

Responses from the visual analogue scales were first recorded as quality scores in the range of 0 to 100. Difference scores were then calculated for each trial and each participant

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TMM.2018.2817070, IEEE Transactions on Multimedia

by subtracting the quality score of the distorted sequence from its corresponding reference. Possible outliers were removed following the procedure in [53]. The numbers of participants discarded are four, three, four and three for Group 1–4 respectively. Difference mean opinion scores (DMOS) were obtained for every trial by taking the mean of the difference scores. The standard errors (SE) of difference scores for each trial were also calculated for subsequent analyses.

Based on participants' responses to the first question, the answers that incorrectly chose distorted sequences with better quality were defined as errors. The correct rate (CR) was then calculated for every trial as the average percentage of non-error responses.

The subjective data, comprised of 384 DMOS, SE, and CR values, have been published alongside the video sequences.

V. Results and Discussion

This section presents the subjective results of the experiments, in the context of the questions proposed in Section I.

As described above, the CR results are based on the responses to the first question in our experiment, which can be considered as a two alternative forced choice (2AFC) task [54]. This enables the CR results to be fitted as a psychometric function with QP as the stimulus level.

The subjective scores of the database were used to evaluate nine popular objective quality metrics, including Peak Signal-to-Noise Ratio (PSNR), Structural Similarity Index (SSIM) [55], multi-scale SSIM (MS-SSIM) [56], Visual Information Fidelity measure (VIF) [57], Visual Signal-to-Noise Ratio (VSNR) [58], Video Quality Metric (VQM) [59], Motion-based Video Integrity Evaluation index (MOVIE) [60], Spatio-temporal Most Apparent Distortion Model (STMAD) [61] and Video Multimethod Assessment Fusion (VMAF) [62]. PSNR, SSIM, MS-SSIM, VIF and VSNR are commonly used image quality metrics, while the remaining three are video quality assessment methods. It is noted that VMAF is a machine learning based video quality metric, which predicts subjective quality by combining multiple existing quality metrics and video feature (including VIF [57], Detail Loss Metric (DLM) [63] and averaged temporal frame difference [62]) using a Support Vector Machine (SVM) regressor [64].

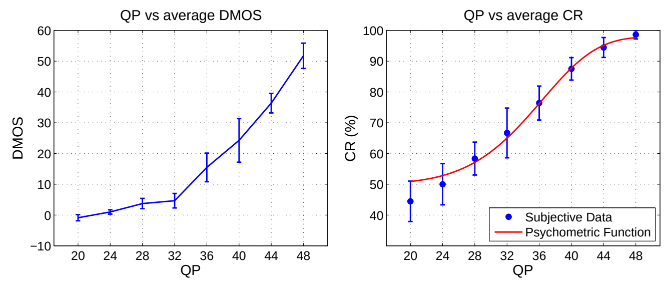
Following the procedure in [19], a logistic fitting function was used to fit the subjective DMOS and objective quality indices based on a weighted least-squares approach. Objective quality metric performance was parameterised using four correlation statistics: the Linear Correlation Coefficient (LCC), the Spearman Rank Order Correlation Coefficient (SROCC), the Outlier Ratio (OR) and the Root Mean Squared Error (RMSE). Definitions of these can be found in [17, 19].

A significance test was also conducted to identify the difference in performance between the objective metrics. The approach in [37, 65] was followed whereby an F-test was performed on the residual between the average DMOS of the BVI-HD database and on the DMOS predicted by the tested objective quality metrics. The predicted DMOS values were obtained based on the best logistic fitting curves.

A. Subjective results on QP range selection (Exp. 1)

Fig. 6.(a) shows the average DMOS over different test QP values for all source sequences in Exp. 1. It is noted that, in most cases, the magnitude of perceptual differences are much more pronounced between QP32 and QP48 than they are between QP20 and QP32.

The CR results plotted against six tested QPs are illustrated in Fig. 6.(b). These produce a tight fit (the deviance and the p-value are 1.986 and 0.937 respectively) to a psychometric function. It can be observed that subjective results on content using the full range of QP values (20-48) correspond to a complete psychometric function. If a correct rate value of 75% is considered as a threhold for detecting compression artefacts [54], it corresponds to QP 36 approximately according to the average psychometric function.



(a) (b)

Fig. 6: Subjective quality results of Exp. 1. (a) and (b) shows the average DMOS and CR over tested QPs respectively. The error bars in both sub-figures represent the standard errors over source sequences.

It is noted that the QP range recommended by the HEVC common test conditions [51] is from 22 to 37, which only covers half of the psychometric function curve in Fig. 6.(b) and misses the visually-lossy part. In order to investigate the subjective quality of HEVC compressed content across a complete psychometric curve, we recommend the use of test QP values of 22, 27, 32, 37, 42, and 47 for future VQA studies. This test range is also suggested for future evaluation of the rate quality performance of HEVC codecs[2].

B. Subjective Quality for HEVC Content (Exp. 2)

The subjective quality scores against various test QPs, in terms of the average DMOS, and the average correct ratios over HEVC compressed content for all source sequences and all participants, are shown in Fig. 7.

It can be observed in Fig. 7.(a) that different characteristics appear between groups with low (22, 27 and 32) and high (37, 42 and 47) QP values. In low QP cases, the average DMOS are below 10, and the compression quality is normally considered as visual lossless or close to. As QP increases, the average DMOS becomes much higher. The average CR in Fig. 7.(b) also show the similar property as Fig. 6.(b), and it almost covers the complete psychometric function curve as expected.

[2]It is noted here that this recommendation is based on the experimental results on HEVC compressed content using constant QP values for all types/levels of frames (QP Offset equals zero). The optimal QP test range may differ if QP values vary among frames.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TMM.2018.2817070, IEEE Transactions on Multimedia

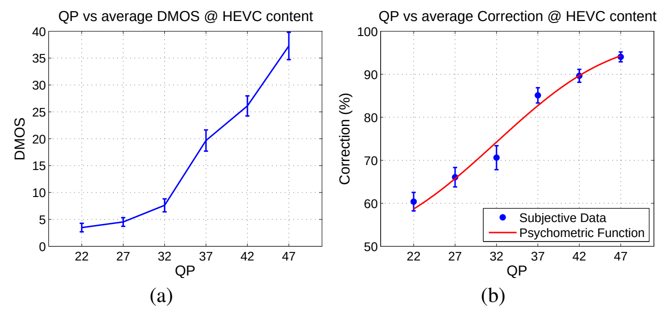


Fig. 7: Subjective quality results for HEVC compressed videos at various QP values. (a) and (b) shows the average DMOS and correct ratios respectively. The error bars in both sub-figures represent the standard errors over source sequences.

C. Subjective quality of synthesised content (Exp. 2)

Fig. 8.(a) shows the proportions of frame blocks that have been synthesised in B_b (the definition is given in the appendix) frames at various QPs and synthesis levels. It is observed that from 30% to 55% of blocks are synthsised in B_b frames, and that this figure is content dependent.

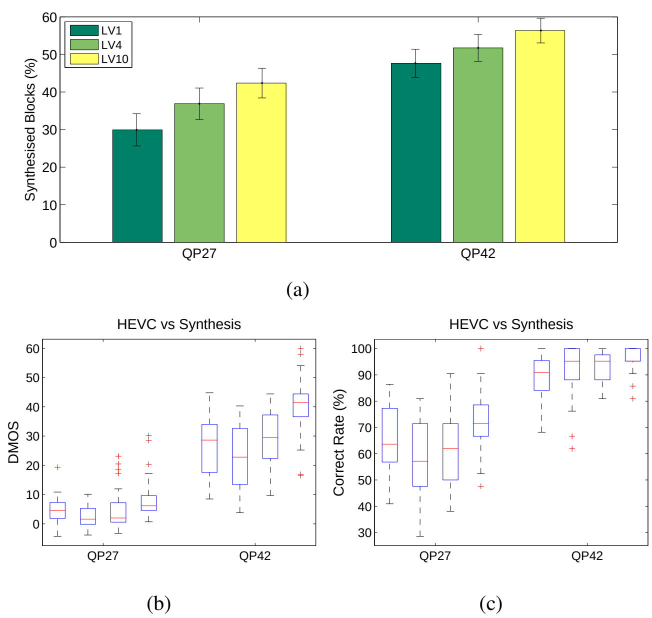


Fig. 8: Results for synthesised content. (a) The percentage of blocks synthesised in B_b frames, where the error bars present the standard errors over content. (b) and (c) show the box plots of average DMOS and correct ratios respectively, for all HEVC and synthesised sequences at QP 27 and 42, where in each four-box group, the first box is for HEVC, and the three other are results for synthesis level 1, 4, and 10.

Fig. 8.(b) presents the average DMOS and CR values for HEVC and various synthesis levels. For all synthesis levels (1, 4, and 10), there is no significant difference in both average DMOS and correct ratio results between HEVC and synthesised content. In some cases, the mean of DMOS and correct ratios for synthesised results are better than the corresponding HEVC compressed videos. This shows the potential of texture synthesis in video compression, particularly in the context of the reduced bit rates achievable.

D. Objective Quality Metric Performance Comparison

The correlation performance of the objective quality metrics against the subjective scores for the database is shown in Fig. 9, TABLE I, and TABLE II. It can be observed that MS-SSIM, VIF, VSNR, VQM, MOVIE, STMAD and VMAF perform better than PSNR and SSIM, with more compact scattering about their fitting curves and higher correlation coefficients. This is also statistically significant based on the F-test results in TABLE II, while the difference is not significant between any two of MS-SSIM, VIF, VSNR, VQM, MOVIE and STMAD. Only VMAF outperforms MS-SSIM, VSNR, MOVIE and STMAD based on the F-test results. For the HEVC subgroup, VMAF offers the highest SROCC value, 0.8014, and VIF performs best for synthesised content. It has been noted that VMAF is a machine learning based quality metric, which combines several existing quality metrics and video features together. Its model parameters were extensively trained on distorted content with compression artifacts [62].

TABLE I: Comparison of correlation statistics for tested video metrics on the BVI-HD video database. The best performer is highlighted in **bold** font for each statistic

Content	All Data				HEVC	Synth
Metric	LCC	SROCC	OR	RMSE	SROCC	SROCC
PSNR	0.6009	0.5923	0.5625	13.8420	0.6194	0.5716
SSIM	0.5748	0.5753	0.5703	14.1642	0.5992	0.5538
MSSSIM	0.6597	0.6615	0.5130	12.1291	0.7158	0.6014
VIF	**0.7658**	**0.7700**	0.4609	11.1051	0.7712	**0.7689**
VSNR	0.7282	0.7362	0.4792	11.8329	0.7408	0.7345
VQM	0.7654	0.7584	0.4401	11.0933	0.7857	0.7261
MOVIE	0.7450	0.7295	0.4401	11.5051	0.7529	0.7030
STMAD	0.7536	0.7541	**0.4141**	11.3898	0.7958	0.7128
VMAF	0.7627	0.7446	0.4323	**10.3530**	**0.8014**	0.6781

It should be also noted that the SROCC values for all tested quality metrics on the whole BVI-HD database are below 0.8, and are even lower (below 0.6) for PSNR and SSIM. This level of performance is lower than achieved by the same metrics on popular databases, such as VQEG FR-TV Phase I [19] and LIVE [37]. This indicates that these metrics perform relatively poorly on HEVC compressed content, something which requires further investigation if such metrics are to be universally accepted. Metric performance is even worse for synthesised content, indicating that significant research is still needed to develop perceptually accurate quality metrics that can deal with synthesis artefacts. It can also be observed for Fig. 9 that the fit of the metrics to the DMOS scores is, in general, significant better for higher quality content.

VI. Conclusions

In this paper, a novel video quality database has been presented for video compression and quality assessment. This database contains 32 original videos that reflect the feature

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TMM.2018.2817070, IEEE Transactions on Multimedia

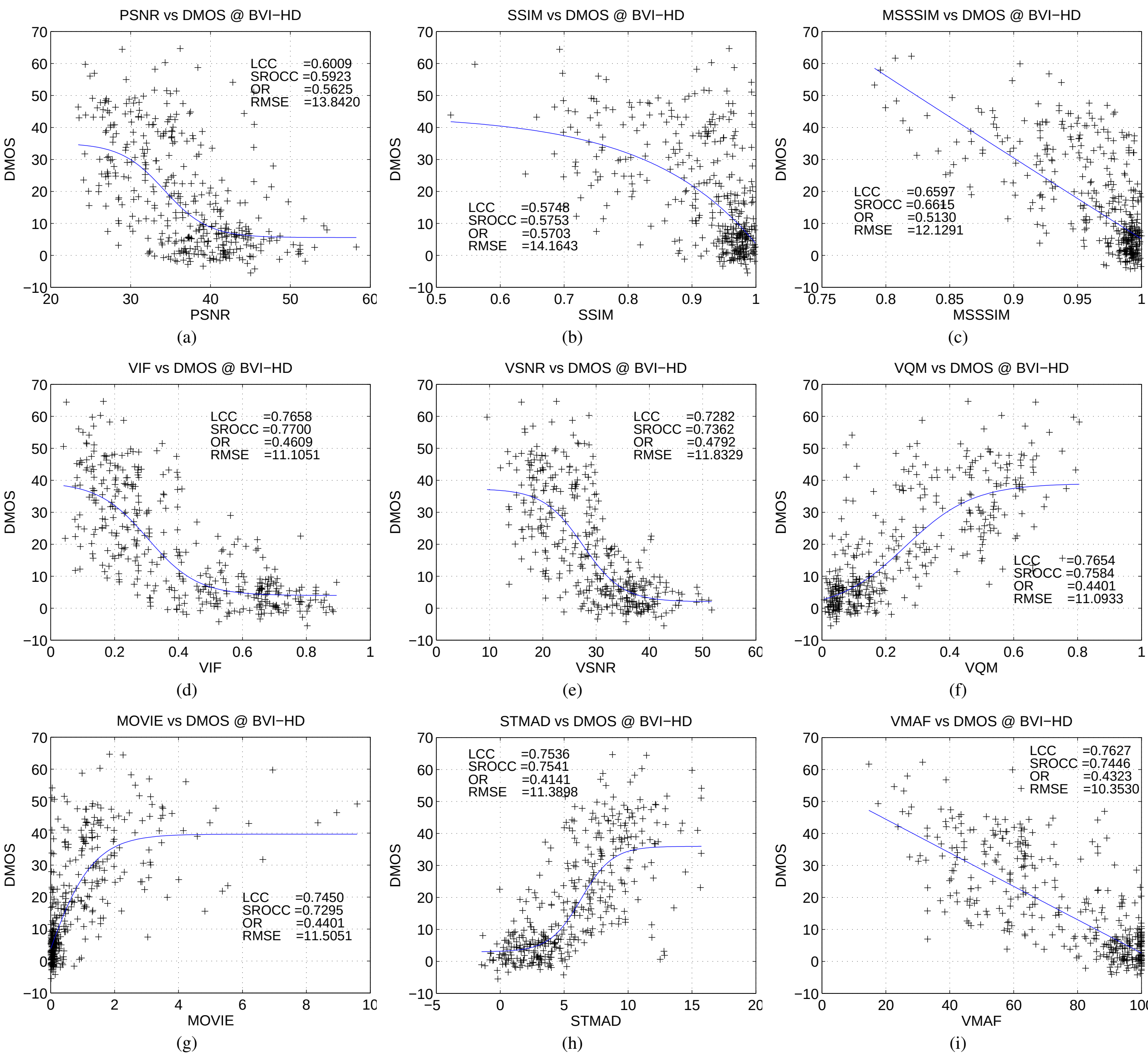


Fig. 9: Scatter plots of subjective DMOS versus the predictions of different quality metrics on the BVI-HD database. The blue curves in the sub-figures represent the best logistic fitting functions.

distributions observed in modern broadcasting content. Each reference sequence is used to generate 12 distorted clips using the current standard video codec, HEVC, and its synthesis integrated version (HEVC-SYNTH).

In order to answer the questions proposed in Section I, two subjective experiments were conducted, the first of which identified an appropriate test range of quantisation levels for HEVC, and recommends the use of wider QP range than that currently in the HEVC common test conditions. The second experiment collected the subjective scores of the database, which were used to test nine popular objective quality assessment models, and to compare the synthesised results with HEVC content. The comparison results show the potential of texture synthesis in video compression, while the performance of quality assessment methods on the BVI-HD database exposes the limitations of current video quality metrics on HEVC, and especially on synthesised content.

The BVI-HD video quality database reported here has been made available online[1] for public testing. Its content diversity and representativeness make it a reliable test database for objective video metric evaluation, and the reference sequences can also be used for video compression performance testing. It should be noted that the number of reference and test sequences in the BVI-HD database is still limited and only full HD resolution has been tested. Future work should focus on using more immersive video formats as test content, including higher dynamic ranges, higher frame rates and higher spatial resolutions.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TMM.2018.2817070, IEEE Transactions on Multimedia

APPENDIX
SYNTHESISED CONTENT GENERATION

Fig. 10 shows a diagram of the algorithm (HEVC-SYNTH) which was used to generate the synthesised video clips in the BVI-HD video quality database. This is a modified version of our synthesis-based encoder in [21].

In order to investigate the subjective quality of synthesised content, and its difference from that of HEVC compressed content, we utilised a one-pass strategy to enable fair comparison. In this approach, following the same GOP structure used in Section III-B, B frames that are inter-predicted only from temporally previous frames (denoted as $\mathrm{B_p}$), and I frames are defined as key frames, and their HEVC reconstructed frames are originally kept in the output video. The three successive bi-directional inter-predicted B frames (denoted as $\mathrm{B_b}$ frames) are processed using the following steps.

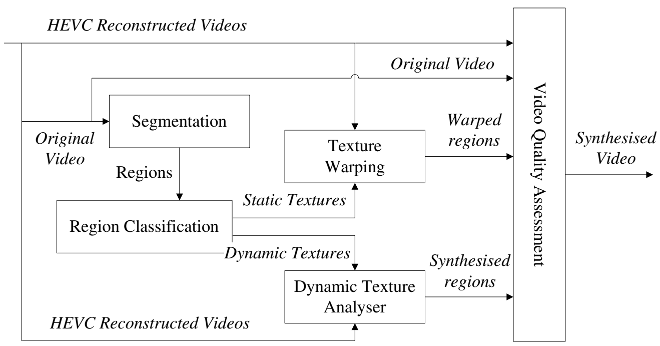


Fig. 10: The approach used to generate synthesised content.

Each $\mathrm{B_b}$ frame is firstly segmented into spatially homogeneous regions, and then classified as static dynamic or non-textured. For non-textured regions, their corresponding HEVC reconstructed content are retained. The static textures in all $\mathrm{B_b}$ frames are firstly warped from their nearest key frames, while the dynamic textures in the first and third (in every three successive $\mathrm{B_b}$ frames, temporal order) $\mathrm{B_b}$ frames are synthesised from their two nearest key frames and the second $\mathrm{B_b}$ frame (possibly with warped content). These warped and synthesised results are then compared with the corresponding original HEVC content in the video quality assessment module, and are conditionally selected based on their quality difference.

In the video quality assessment module, a recently developed video quality metric, PVM, is employed [65, 66], which is an enhanced version of AVM (the artefact-based video metric) that is used in [21]. Every 64×64 block of warped/synthesised content and their original HEVC counterpart are compared using PVM predicted DMOS, which is converted from original PVM indices using a non-linear fitting curve. This fitting function is defined in [19], and its parameters were obtained based on the correlation between PVM quality indices and subjective results in the LIVE video database. Only when the difference between the PVM predicted DMOS between a warped/synthesised block and the corresponding HEVC content is lower or equal to a certain threshold, will the warped/synthesised block replace the HEVC content. Otherwise the HEVC compressed content will remain in the output. The threshold was configured as 1, 4, and 10, where larger value allows more synthesis.

The texture segmentation, classification, warping and synthesis methods used in this algorithm, and the definition of PVM can be found in [21, 65].

REFERENCES

[1] M. Armstrong, D. Flynn, M. Hammond, S. Jolly, and R. Salmon, "High frame-rate television," *BBC Research White Paper 169*, 2008.
[2] Recommendation ITU-R BT.2020, "Parameter values for ultra-high definition television systems for production and international programme exchange," ITU-R, Tech. Rep., 2012.
[3] T. Tokimoto, K. Sato, S. Suyama, and H. Yamamoto, "High-frame-rate LED display with pulse-width modulation by use of nonlinear clock," in *Consumer Electronics (GCCE), 2013 IEEE 2nd Global Conference on*. IEEE, 2013, pp. 83–84.
[4] R. Atkins, "Method and apparatus in various embodiments for hdr implementation in display devices," Patent, Jun. 25, 2009, US Patent App. 12/491,857.
[5] H. Hadizadeh and I. V. Bajic, "Full-reference objective quality assessment of tone-mapped images," *IEEE Trans. on Multimedia*, 2017.
[6] A. Mackin, F. Zhang, and D. R. Bull, "A study of subjective video quality at various frame rates," in *Proc. IEEE Int Conf. on Image Processing*. IEEE, 2015.
[7] F. Zhang, A. Mackin, and D. R. Bull, "A new frame rate dependant quality model based on temporal wavelet decomposition and spatiotemporal pooling," in *Proc. IEEE Int Conf. on Image Processing*. IEEE, 2017.
[8] R. Mantiuk, K. J. Kim, A. G. Rempel, and W. Heidrich, "HDR-VDP-2: A calibrated visual metric for visibility and quality predictions in all luminance conditions," in *ACM Transactions on Graphics (TOG)*, vol. 30, no. 4. ACM, 2011, p. 40.
[9] J. G. Apostolopoulos, P. A. Chou, B. Culbertson, T. Kalker, M. D. Trott, and S. Wee, "The road to immersive communication," *Proceedings of the IEEE*, vol. 100, no. 4, pp. 974–990, 2012.
[10] Y. Dong, M. T. Pourazad, and P. Nasiopoulos, "Human visual system-based saliency detection for high dynamic range content," *IEEE Trans. on Multimedia*, vol. 18, no. 4, pp. 549–562, 2016.
[11] F. Ribeiro, J. de Oliveira, A. Ciancio, E. da Silva, C. Estrada, L. Tavares, J. Gois, M. Martelotte, and A. Said, "Quality of experience in a stereoscopic multiview environment," *IEEE Trans. on Multimedia*, 2017.
[12] Y. Chen, X. Zhao, L. Zhang, and J.-W. Kang, "Multiview and 3D video compression using neighboring block based disparity vectors," *IEEE Trans. on Multimedia*, vol. 18, no. 4, pp. 576–589, 2016.
[13] G. J. Sullivan, J. R. Ohm, W. J. Han, and T. Wiegand, "Overview of the high efficiency video coding (HEVC) standard," *IEEE Trans. on Circuits and Systems for Video Technology*, vol. 22, no. 12, pp. 1649–1668, 2012.
[14] G. J. Sullivan, J. M. Boyce, Y. Chen, J.-R. Ohm, C. A. Segall, and A. Vetro, "Standardized extensions of high efficiency video coding (HEVC)," *IEEE Journal of selected topics in Signal Processing*, vol. 7, no. 6, pp. 1001–1016, 2013.
[15] E. Upenik, M. Rerabek, and T. Ebrahimi, "On the performance of objective metrics for omnidirectional visual content," in *9th International Conference on Quality of Multimedia Experience (QoMEX 2017)*, 2017.
[16] A. F. Silva, M. C. Q. Farias, and J. A. Redi, "Perceptual annoyance models for videos with combinations of spatial and temporal artifacts," *IEEE Trans. on Multimedia*, vol. 18, no. 12, pp. 2446–2456, 2016.
[17] D. R. Bull, *Communicating pictures: a course in image and Video Coding*. Academic Press, 2014.
[18] G. Sullivan, J.-R. Ohm, and V. Baroncini, "Preliminary joint call for evidence on video compression with capability beyond hevc," Joint Video Exploration Team (JVET) of ITU-T VCEG and ISO/IEC MPEG, Tech. Rep., 2017.
[19] Video Quality Experts Group, "Final report from the video quality experts group on the validation of objective metrics for video quality assessment." VQEG, Tech. Rep., 2000. [Online]. Available: http://www.its.bldrdoc.gov/vqeg/projects/frtv_phaseI
[20] S. Winkler, "Analysis of public image and video database for quality assessment," *IEEE Journal of Selected Topics in Signal Processing*, vol. 6, no. 6, pp. 1–10, 2012.
[21] F. Zhang and D. R. Bull, "A parametric framework for video compression using region-based texture models," *IEEE Journal of Selected Topics in Signal Processing*, vol. 5, no. 7, pp. 1378–1392, 2011.
[22] P. Ndjiki-Nya, T. Hinz, and T. Wiegand, "Generic and robust video

TABLE II: F-test results for video metrics at 95% confidence level. The number here indicates the F-test results for the whole BVI-HD database. “1” suggests that the metric in the row is superior to that in the column (“-1” if the oppsite is true), while “0” indicates that there is no significant difference between them. The degrees of freedom for BVI-HD video quality database is 383.

Metric	PSNR	SSIM	MS-SSIM	VIF	VSNR	VQM	MOVIE	STMAD	VMAF
PSNR	-	0	-1	-1	-1	-1	-1	-1	-1
SSIM	0	-	-1	-1	-1	-1	-1	-1	-1
MSSSIM	1	1	-	0	0	0	0	0	-1
VIF	1	1	0	-	0	0	0	0	0
VSNR	1	1	0	0	-	0	0	0	-1
VQM	1	1	0	0	0	-	0	0	0
MOVIE	1	1	0	0	0	0	-	0	-1
STMAD	1	1	0	0	0	0	0	-	-1
VMAF	1	1	1	0	1	0	1	1	-

coding with texture analysis and synthesis,” in *Proc. IEEE Int Conf. on Multimedia and Expo*. IEEE, 2007, pp. 1447–1450.

[23] M. Bosch, F. Zhu, and E. J. Delp, “Segmentation-based video compression using texture and motion models,” *IEEE Journal of Selected Topics in Signal Processing*, vol. 5, no. 7, pp. 1277–1281, 2011.

[24] J. Balle, A. Stojanovic, and J. R. Ohm, “Models for static and dynamic texture synthesis in image and video compression,” *IEEE Journal of Selected Topics in Signal Processing*, vol. 5, no. 7, pp. 1353–1365, 2011.

[25] U. S. Thakur and O. Chubach, “Texture analysis and synthesis using steerable pyramid decomposition for video coding,” in *International Conference on Systems, Signals and Image Processing*. IEEE, 2015, pp. 204–207.

[26] C.-H. Yeh, T.-Y. Tseng, C.-W. Lee, and C.-Y. Lin, “Predictive texture synthesis-based intra coding scheme for advanced video coding,” *IEEE Trans. on Multimedia*, vol. 17, no. 9, pp. 1508–1514, 2015.

[27] B. Butterworth, “History of the‘BBC Redux’project,” *BBC Internet Blog*, 2013.

[28] F. Mercer Moss, F. Zhang, R. Baddeley, and D. R. Bull, “What is on TV: A large scale quantitative characterisation of modern broadcast video content,” in *Proc. IEEE Int Conf. on Image Processing*. IEEE, 2016.

[29] Recommendation ITU-R BT.500-11, *Methodology for the subjective assessment of the quality of television pictures*, Std., 2002.

[30] F. Mercer Moss, K. Wang, F. Zhang, R. Baddeley, and D. R. Bull, “On the optimal presentation duration for subjective video quality assessment,” *IEEE Trans. on Circuits and Systems for Video Technology*, vol. 26, no. 11, pp. 1977–1987, 2016.

[31] F. Mercer Moss, C.-T. Yeh, F. Zhang, R. Baddeley, and D. R. Bull, “Support for reduced presentation durations in subjective video quality assessment,” *Signal Processing: Image Communication*, vol. 48, pp. 38–49, 2016.

[32] J. E. Cutting, K. L. Brunick, J. E. DeLong, C. Iricinschi, and A. Candan, “Quicker, faster, darker: Changes in hollywood film over 75 years,” *i-Perception*, vol. 2, no. 6, pp. 569–576, 2011.

[33] P. Fröhlich, S. Egger, R. Schatz, M. Mühlegger, K. Masuch, and B. Gardlo, “Qoe in 10 seconds: Are short video clip lengths sufficient for quality of experience assessment?” in *Quality of Multimedia Experience (QoMEX), 2012 Fourth International Workshop on*. IEEE, 2012, pp. 242–247.

[34] B. W. Tatler, R. J. Baddeley, and I. D. Gilchrist, “Visual correlates of fixation selection: effects of scale and time,” *Vision Research*, vol. 45, no. 5, pp. 643–659, 2005.

[35] F. J. Mercer Moss, R. Baddeley, and N. Canagarajah, “Eye movements to natural images as a function of sex and personality,” *PloS one*, vol. 7, no. 11, p. e47870, 2012.

[36] P.910, “Recommendation ITU-T P.910: Subjective video quality assessment methods for multimedia applications,” ITU-T, Tech. Rep., 1999.

[37] K. Seshadrinathan, R. Soundararajan, A. C. Bovik, and L. K. Cormack, “Study of subjective and objective quality assessment of video,” *IEEE Trans. on Image Processing*, vol. 19, pp. 335–350, 2010.

[38] Video Quality Experts Group, “Report on the validation of video quality models for high definition video content,” VQEG, Tech. Rep., 2010. [Online]. Available: http://www.its.bldrdoc.gov/vqeg/projects/hdtv/hdtv.aspx

[39] M. H. Pinson, “The consumer digital video library [best of the web],” *IEEE Signal Processing Magazine*, vol. 30, no. 4, pp. 172–174, 2013.

[40] S. Pechard, R. Pepion, and P. L. Callet, “Suitable methodology in subjective video quality assessment a resolution dependent paradigm,” in *Proc. International Workshop on Image Media Quality and its Applications (IMQA)*, 2008.

[41] F. Zhang, S. Li, L. Ma, Y. C. Wong, and K. N. Ngan, “IVP subjective quality video database,” 2009. [Online]. Available: http://ivp.ee.cuhk.edu.hk/research/database/subjective/

[42] F. D. Simone, “EPFL-PoliMI video quality assessment database,” 2009. [Online]. Available: http://vqa.como.polimi.it

[43] J. S. Lee, F. De Simone, and T. Ebrahimi, “Subjective quality evaluation via paired comparison: application to scalable video coding,” *IEEE Trans. on Multimedia*, vol. 13, no. 5, pp. 882–893, 2011.

[44] Y.-F. Ou, T. Liu, Z. Zhao, Z. Ma, and Y. Wang, “Modeling the impact of frame rate on perceptual quality of video,” in *Proc. IEEE Int Conf. on Image Processing*. IEEE, 2008, pp. 689–692.

[45] P. V. Vu and D. M. Chandler, “ViS3: an algorithm for video quality assessment via analysis of spatial and spatiotemporal slices,” *Journal of Electronic Imaging*, vol. 23, no. 1, 2014.

[46] M. A. Papadopoulos, F. Zhang, D. Agrafiotis, and D. R. Bull, “A video texture database for perceptual compression and quality assessment,” in *Proc. IEEE Int Conf. on Image Processing*. IEEE, 2015.

[47] Y. Zhu, L. Song, R. Xie, and W. Zhang, “SJTU 4K video subjective quality dataset for content adaptive bit rate estimation without encoding,” in *Broadband Multimedia Systems and Broadcasting (BMSB), 2016 IEEE International Symposium on*. IEEE, 2016, pp. 1–4.

[48] M. Afonso, A. Katsenou, F. Zhang, D. Agrafiotis, and D. Bull, “Video texture analysis based on HEVC encoding statistics,” in *Picture Coding Symposium (PCS)*, 2016.

[49] M. Cheon and J.-S. Lee, “Subjective and objective quality assessment of compressed 4K UHD videos for immersive experience,” *IEEE Trans. on Circuits and Systems for Video Technology*, 2017.

[50] G. W. Corder and D. I. Foreman, *Nonparametric statistics: A step-by-step approach*. John Wiley & Sons, 2014.

[51] F. Bossen, “Common test conditions and software reference configurations,” no. JCTVC-L1100. ITU-T, ISO/IEC, 2013.

[52] K. Sharman and K. Suehring, “Common test conditions,” no. JCTVC-X1100. ITU-T, ISO/IEC, 2016.

[53] Recommendation ITU-R BT.1788, *Methodology for the subjective assessment of video quality in multimedia applications*, ITU-R Std., 2007.

[54] G. A. Gescheider, *Psychophysics: the fundamentals*. Psychology Press, 2013.

[55] Z. Wang, A. Bovik, H. Sheikh, and E. Simoncelli, “Image quality assessment: from error visibility to structural similarity,” *IEEE Trans. on Image Processing*, vol. 13, pp. 600–612, 2004.

[56] Z. Wang, E. P. Simoncelli, and A. C. Bovik, “Multi-scale structural similarity for image quality assessment,” in *Proc. Asilomar Conference on Signals, Systems and Computers*, vol. 2. IEEE, 2003, p. 1398.

[57] H. R. Sheikh, A. C. Bovik, and G. de Veciana, “An information fidelity criterion for image quality assessment using natural scene statistics,” *IEEE Trans. on Image Processing*, vol. 14, pp. 2117–2128, 2005.

[58] D. Chandler and S. Hemami, “VSNR: A wavelet-based visual signal-to-noise ratio for natural images,” *IEEE Trans. on Image Processing*, vol. 16, no. 9, pp. 2284–2298, 2007.

[59] M. H. Pinson and S. Wolf, "A new standardized method for objectively measuring video quality," *IEEE Trans. on Broadcasting*, vol. 50, pp. 312–322, 2004.

[60] K. Seshadrinathan and A. C. Bovik, "Motion tuned spatio-temporal quailty assessment of natural videos," *IEEE Trans. on Image Processing*, vol. 19, pp. 335–350, 2010.

[61] P. V. Vu, C. T. Vu, and D. M. Chandler, "A spatiotemporal most-apparent-distortion model for video quality assessment," in *Proc. IEEE Int Conf. on Image Processing*. IEEE, 2011, pp. 2505–2508.

[62] Z. Li, A. Aaron, I. Katsavounidis, A. Moorthy, and M. Manohara, "Toward a practical perceptual video quality metric," *The Netflix Tech Blog*, vol. 6, 2016.

[63] S. Li, F. Zhang, L. Ma, and K. H. Ngan, "Image quality assessment by separately evaluating detail losses and additive impairments," *IEEE Transactions on Multimedia*, vol. 13, no. 5, pp. 935–949, 2011.

[64] C. Cortes and V. Vapnik, "Support-vector networks," *Machine learning*, vol. 20, no. 3, pp. 273–297, 1995.

[65] F. Zhang and D. R. Bull, "A perception-based hybrid model for video quality assessment," *IEEE Trans. on Circuits and Systems for Video Technology*, vol. 26, no. 6, pp. 1017–1028, 2016.

[66] F. Zhang and D. Bull, "Quality assessment method for perceptual video compression," in *Proc. IEEE Int Conf. on Image Processing*, 2013, pp. 39–43.

Fan Zhang (M'12) received the B.Sc. and M.Sc. degrees from Shanghai Jiao Tong University (2005 and 2008 respectively), and his Ph.D from the University of Bristol (2012). He is currently working as a Research Assistant in the Visual Information Laboratory, Department of Electrical and Electronic Engineering, University of Bristol, on projects related to perceptual video compression. His research interests include perceptual video compression, video quality assessment and immersive video formats including HDR and HFR.

Felix Mercer Moss (M'16) received his B.Sc from the University of Leeds (2008) before obtaining his M.Res (2009) and Ph.D (2013) from the departments of Experimental Psychology and Computer Science at the University of Bristol, respectively. He is currently working as a research assistant in the Visual Information Laboratory in the department of Electrical and Electronic Engineering, University of Bristol, working on projects related to video quality assessment. His research interests include the subjective video quality assessment, image evaluation, individual differences in eye movement control and decision making.

Roland Baddeley obtained his BSc from the University of Sussex, and Phd from the University of Stirling. His previous roles include positions at Cambridge Physiology, Oxford Psychology and Physiology and Sussex Psychology departments. He is currently a reader in computational neuroscience in the school of Experimental Psychology at University of Bristol. He has published over 50 academic papers and his research interests include the statistics on natural images, eye movement control, camouflage, decision making, and the perception of texture.

David R. Bull (M'94-SM'07-F'12) received the B.Sc. degree from the University of Exeter, Exeter, U.K., in 1980; the M.Sc. degree from University of Manchester, Manchester, U.K., in 1983; and the Ph.D. degree from the University of Cardiff, Cardiff, U.K., in 1988.

Dr Bull has previously been a Systems Engineer with Rolls Royce, Bristol, U.K. and a Lecturer at the University of Wales, Cardiff, U.K. He joined the University of Bristol in 1993 and is currently its Chair of Signal Processing and Director of its Bristol Vision Institute. In 2001, he co-founded a university spin-off company, ProVision Communication Technologies Ltd., specializing in wireless video technology. He has authored over 450 papers on the topics of image and video communications and analysis for wireless, Internet and broadcast applications, together with numerous patents, several of which have been exploited commercially. He has received two IEE Premium awards for his work. He is the author of three books, and has delivered numerous invited/keynote lectures and tutorials. Dr. Bull is a fellow of the Institution of Engineering and Technology.