Signal Processing: Image Communication 78 (2019) 152–158

Contents lists available at ScienceDirect

Signal Processing: Image Communication

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

A multi-metric approach for block-level video quality assessment ☆

Miltiadis Alexios Papadopoulos, Angeliki V. Katsenou *, Dimitris Agrafiotis, David R. Bull

Visual Information Lab, Department of Electrical and Electronic Engineering, University of Bristol, Bristol, UK

ARTICLE INFO

Keywords:
Video quality assessment
Content features
Fusion of metrics
Block-level
RDO

ABSTRACT

Developing an objective video quality metric that accurately estimates perceived video quality is challenging. Developing a metric that can additionally be embedded in the rate distortion optimization process of a video codec can be even harder given that decisions have to be made locally. In this paper, we present a method for combining a number of existing state of the art objective video quality metrics at the coding block level by employing a fusion of local content features for deciding how to best utilize the chosen metrics. Our results indicate promising performance in terms of the correlation of the developed locally-acting quality metric with the overall perceived quality of the video.

1. Introduction

Although video quality has been traditionally evaluated using Mean Squared Error (MSE), it is already known that it does not linearly correlate with the perceived quality due to the human visual system properties that are not captured by it [1]. The most reliable method to assess the quality of the compressed videos is through the subjective assessment of the perceived quality. This, however, for a real-time system is impractical due to the time constraints imposed. As a solution, many different objective quality metrics that purport to correlate well with perceived quality have been proposed. However, the performance of these metrics varies widely on different video content [2].

The literature is rich in quality metrics which claim better correlation to perceptual quality than MSE. These metrics were either initially designed for images, such as the Structural Similarity Index (SSIM) [1], Peak Signal to Noise Ratio based on HVS (PSNRHVSM) [3], Multi-Scale SSIM (MS-SSIM) [4], Visual Information Fidelity (VIF) [5], Feature Similarity Index (FSIM) [6]; or for video, such as Perception-based Video Metric (PVM) [7], Motion-based Video Integrity Evaluation (MOVIE) index [8], or Video Quality Metric (VQM) [9]. Although most of the aforementioned metrics correlate better with perceived quality than PSNR [10] for compressed video, they lack the capability of operating as an integral part of the RDO process, either because they are highly complex (e.g. MOVIE) or because they do not offer the additive property; the measured quality of a region is not equal to the sum of measured quality of its parts. RDO addresses this problem by utilizing the SAD and SATD metrics that offer such a property up to the CTU level; RDO optimizations are performed on each level of block segmentation. However, this is limited to the size of the CTU. Several CTUs are never assessed together and therefore their collective score is never calculated for the purposes of RDO. Our work is a method for assessing the overall quality at a different segmentation level, as our method segments the CTUs based on their content characteristics. Some of the metrics above have been tested within an RDO framework. SSIM is typically an example of such an attempt (e.g. [11–13]) which has been applied to RDO [14] and quantization [15]. SSIM is also an example of a metric that does not offer the additive property, rendering it difficult for use by the RDO process. It is also important to note that improving PSNR [16] by adapting it to subjective quality evaluation scores has received extensive research.

Choosing amongst all these metrics is a challenge by itself as they each offer different levels of performance for different content. One way to address metric selection is through fusion of several metrics using machine learning techniques. VMAF [17] is a good example of a practical quality metric that fuses VIF [5], DLM [18] and motion information (i.e. frame differencing). Being trained on a large varied dataset, VMAF shows higher correlation to subjective quality compared to other objective quality metrics. However, it evaluates the overall frame quality, which is not ideal in an RDO environment where block-level quality estimation is required.

Motivated by the above, this work introduces a block-level fusion of objective metrics for video quality assessment (BVQA). BVQA is a result of fusing state of the art objective metrics based on their spatio-temporal content at a block level. A diagrammatic outline of the proposed method to develop BVQA models is depicted in Fig. 1. First a small scale study is performed on a set of best-performing objective metrics. Then, based on this, a content analysis takes place. In

☆ No author associated with this paper has disclosed any potential or pertinent conflicts which may be perceived to have impending conflict with this work. For full disclosure statements refer to https://doi.org/10.1016/j.image.2019.06.009.

* Correspondence to: Office 1.23, 1 Cathedral Square, B1 5DD, Bristol, UK.
E-mail address: angeliki.katsenou@bristol.ac.uk (A.V. Katsenou).

https://doi.org/10.1016/j.image.2019.06.009
Received 2 September 2018; Received in revised form 23 June 2019; Accepted 24 June 2019
Available online 28 June 2019

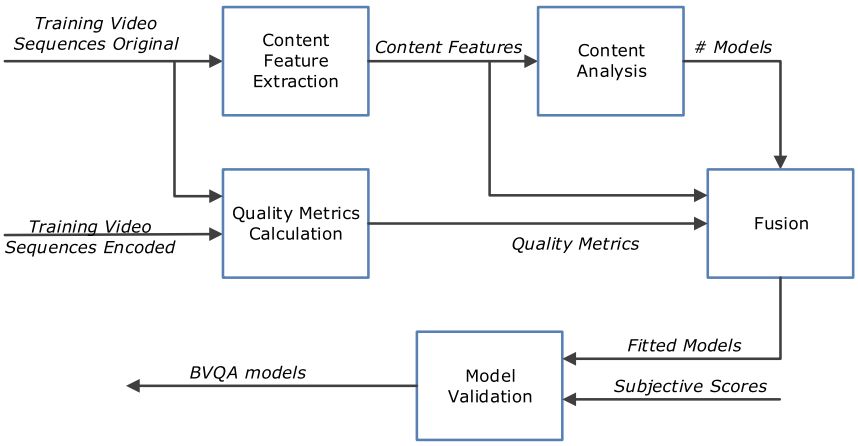


Fig. 1. Diagrammatic outline of the proposed method to develop the BVQA models.

Table 1
Linear & Rank correlation of DMOS & SOA metrics for BVI_Texture data set and relative average complexity.

		FSIM	MSSSIM	PSNR	PSNR-HVS	PSNR-HVSM	SSIM	VIFP
Blk	Lin	**.629**	.623	.257	.280	.365	.474	.507
	Rnk	.605	**.651**	.443	.499	.609	.526	.522
Frm	Lin	**.736**	.726	.325	.355	.453	.544	.615
	Rnk	.697	**.804**	.479	.523	.640	.658	.600
Seq	Lin	**.746**	.743	.327	.357	.457	.551	.621
	Rnk	.709	**.815**	.479	.525	.639	.665	.602
Rel. Cmplx		3.00	1.82	**1**	1.36	1.36	1.55	2.18

Table 2
Cluster centroids in the three content feature dimensions.

Cluster	EDGE_ENT	SI	TI
K0	.126	.141	.138
K1	.251	.456	.061
K2	.392	.585	.048
K3	.443	.576	.270
K4	.059	.050	.031
K5	.164	.288	.039
K6	.170	.254	.328

particular, content-based clustering of the video blocks is performed to group blocks with similar content features and quality. Considering this grouping of content and quality, different block-level quality prediction models are developed. The aim here is to identify the fusion of metrics that performs best in specific scenarios, as some metrics might perform better at relatively static scenes compared to others. BVQA does not aim to provide the equation that accurately describes subjective quality based on simple objective metrics but rather attempts to estimate it. Moreover, three different categories of models of different levels of complexity are examined. All three categories consider the content features in the fusion of metrics into models and take advantage of the fact that the correlation of the objective metrics to the perceptual quality depends on the content features. To the best of our knowledge, this is the first time a content-driven fusion of objective metrics at a block level has been proposed in the literature.

The rest of the paper is arranged as follows: in Section 2 we do a small study on the performance of state-of-the-art (SOA) metrics at a block level. In Section 3, we perform a content analysis of the different blocks of the considered video sequences with the aim of identifying groups of content that have similar quality performance. Based on this, we introduce a content-driven multi-metric fusion approach at a block level in Section 4. Finally, in Section 5 conclusions are drawn.

2. Quality evaluation at the block level

In recent years, the assumption of optimizing on short video clips (at a "chunk" or "shot" level) has been adopted either with respect to the trade-off between streaming performance and coding efficiency [19–22] or because of the trade-off of the presentation duration and the scoring for subjective quality purposes [23]. If we assume that for short-duration videos (up to 5 s) the spatial and temporal characteristics are consistent (within one shot), then the perceived quality after compression is expected to be effectively the same across all frames. Moreover, if we consider sequences with no apparent viewing pattern, the foveation effects are omitted and the perceived video quality is not expected to change dramatically within a frame. To this end, the dataset employed here is one with sequences of one shot and without an obvious focal point. This is the BVI_Texture dataset [24] that contains 20 full high definition (HD) video sequences at 60fps and is annotated with differential mean opinion scores (DMOS). This specific dataset has been selected for two important reasons: firstly, it satisfies the criterion for spatial and temporal homogeneity that allows the extrapolation of the content evaluation scores to the block level. Secondly, the subjective tests were performed in our lab and the raw subjective scores were available. There do not exist many datasets at HD resolution at 60 fps with no apparent viewing pattern that are also providing subjective assessment scores.

The sequences were encoded using the HEVC HM 16.2 (CTC Low Delay mode) at four different compression levels (different quantization levels) and then we computed the value of seven objective quality assessment metrics for each block as reported in Table 1. We would like to note that we did not use metrics that have shown better correlation to perceptual quality like PVM, MOVIE, and VQM due to their high complexity and their design to operate at a frame level. The sequences were divided into 64×64 pixel blocks, so that the size and positioning coincides with the block partitioning of HEVC HM (i.e. CTUs). This created a total of 11.52×10^6 paired data points (i.e. 20 sequences of four different compression levels, 300 frames per sequence, 480 blocks per frame) of subjective quality score and objective quality metric value pairs at a block level.

Looking into the raw data pairs prior to any processing, we report the absolute values of the linear (Lin) and rank (Rnk) correlation coefficients in Table 1 when the metrics are calculated at a per block (Blk), per frame (Frm) and per sequence (Seq) level. The metrics have been calculated at a block-level and the fusion occurs at a block level. The correlation is computed at this level. Then based on the segmentation of the frame (depending on where each block belongs to), a weighted average is computed per frame (weighted average because of the different number of blocks per content class). At this level the

Table 3
Correlation between DMOS and metrics per cluster.

Cl.	% of inst.	FSIM		MSSSIM		PSNR		PSNR-HVS		PSNR-HVSM		SSIM		VIFP	
		Lin	Rnk	Lin	Rnk	Lin	Rnk	Lin	Rnk	Lin	Rnk	Lin	Rnk	Lin	Rnk
K0	17%	.734	.651	**.797**	.810	.719	.748	.731	.775	.752	**.811**	.768	.715	.736	.753
K1	12%	**.844**	.884	.776	.860	.727	.850	.767	.881	.781	**.898**	.797	.856	.796	.869
K2	8%	**.918**	**.913**	.824	.883	.847	.898	.855	.906	.840	.910	.899	.896	.859	.897
K3	7%	.828	.830	.781	.816	.726	**.843**	.706	.832	.698	.831	**.838**	.819	.763	.840
K4	36%	.588	.481	**.713**	.606	.399	.619	.397	.613	.411	**.636**	.588	.516	.290	.127
K5	15%	**.703**	.686	.698	**.773**	.392	.541	.428	.586	.522	.691	.472	.599	.595	.615
K6	6%	.565	.579	.609	.737	.625	.715	.657	.771	**.689**	**.813**	.557	.658	.648	.678

frame correlation to subjective scores is computed. Finally, all frame scores are averaged over the length of frames and the sequence level correlation over all sequences is computed. In Table 1, we observe that in most cases the best performing metrics are FSIM and MS-SSIM. Another important observation is that the correlation coefficients increase as we move from the block level to the frame and then to the sequence level. This is expected because of the different distributions of the metric values at the different spatial levels.

Furthermore, in order to give an idea of the complexity in terms of execution times, in the bottom row of Table 1, we report the relative average complexity of the metrics as ratios of the average execution time over the minimum average execution time. As can bee seen, PSNR requires the lowest execution time on average. On the other hand, FSIM, which is one the most well performing metrics in this table, is concurrently the most expensive in terms of execution time.

3. Content analysis

In this section, we study the quality performance of video blocks with similar content features. Therefore, we propose the clustering of blocks into groups according to their content. As a first step, we calculate three spatio-temporal features for all blocks of the considered sequences. These help identify content characteristics. The selected features are edge entropy (EDGE_ENT), spatial information (SI) and temporal information (TI) that are also used in the ITU-T P.910 recommendation [25]. SI is based on the Sobel filter and expresses the temporal maximum of the standard deviation of luminance over the filtered frame. TI represents the temporal maximum of the standard deviation of spatial differences of adjacent frames. To determine the edge entropy of a block, we first search for regular and homo-directional edges in the scene using the directional edge entropy approach [26,27]. First, a Sobel filter is applied to determine the horizontal and vertical gradients and after determining the direction of edges in every block, we calculate the 73 bins histogram for the values $-180°$ to $180°$, equivalent to a resolution of $5°$ per bin. The edge entropy is given by:

$$\text{EDGE\_ENT} = -\sum_{i=1}^{73}(b_i \cdot log10(b_i)), \quad (1)$$

where b_i is the number of observations for the bin i. The data collected during feature extraction, are then randomized and $1/10$-th of them are selected to be used for k-means clustering (due to software and memory limitations). To avoid cluster biasing, especially in the case of the TI, all features were normalized in the range $[0, 1]$.

Then, in order to select the optimal number of clusters, we employed the Expectation Maximization (EM) algorithm [28] and the elbow method [29]. According to the latter method, we check the ratio of the within class to across classes distortion:

$$D_R = \frac{\sum_{k=1}^{N_{\max}}(I_k - c_k)^2}{\sum_{k,m=1,m\neq k}^{N_{\max}}(I_k - c_m)^2} \quad (2)$$

where I is the data point with coordinates (EDGE_ENT, SI, TI), c_k is the centroid of the k_{th} cluster and $N_{\max}$ is the maximum number of clusters to be considered. During the elbow method application, k-means clustering was applied following a five-fold centroid initialization. Fig. 2(a) depicts D_R for the different number of clusters tested. By inspecting this figure, we observe that the distortion ratio converges at seven clusters. The same number of clusters is suggested by EM clustering when applied on the three content attributes and the seven SOA metrics. Therefore, we conclude that seven clusters is the optimal number of clusters for our data. Fig. 2(b) shows the clustered blocks in the content feature space and Table 2 reports the corresponding cluster centroids. We note that the clustering approach has grouped together regions that feature similar content characteristics. Indeed, K1, K2 and K3 clusters express high SI and EDGE_ENT values, indicating mostly static textures, whereas K0, K4 and K5 clusters are populated by data points that belong to dynamic textures of high TI.

Table 4
Uniformity and coverage of the three content attributes.

No	Sequence	U–Uniformity				T	Score
		EDGE_ENT	SI	TI	Mean	Cover.	U_{mean}·T
1	BallUnderWater	.110	.374	.212	.232	.373	.087
2	Bookcase	.398	.674	.761	.611	.539	.330
3	BriscKBushesStatic	.655	.673	.299	.543	.395	.214
4	BricksLeaves	.583	.588	.385	.518	.446	.231
5	BubblesClear	.049	.462	.338	.283	.213	.060
6	CalmingWater	.389	.522	.484	.465	.389	.181
7	CarpetCircleFast	.195	.217	.236	.216	.256	.055
8	CarpetPanAverage	.286	.239	.249	.258	.278	.072
9	CarpetSlowTrans	.312	.417	.339	.356	.286	.102
10	DropsOnWater	.308	.497	.483	.429	.468	.201
11	Flowers2	.561	.703	.065	.443	.369	.164
12	LampLeaves	.642	.748	.436	.608	.436	.265
13	PaintingTilting	.623	.553	.360	.512	.440	.225
14	PaperStatic	.274	.209	.000	.161	.149	.024
15	PlasmaFree	.415	.543	.362	.440	.586	.258
16	PondDragonflies	.546	.692	.008	.415	.300	.124
17	SmokeClear	.127	.094	.139	.120	.247	.030
18	Sparkler	.415	.485	.473	.458	.556	.255
19	Squirrel	.561	.667	.005	.411	.303	.125
20	TreeWills	.706	.751	.042	.500	.390	.195
	ALL	.643	.779	.429	.617	.741	.457
	Training set	.627	.712	.458	.599	.729	.436
	Testing set	.614	.805	.349	.590	.563	.332

Table 3 lists the linear and rank correlation coefficients for the clustered blocks. It is clear that the metrics perform differently for blocks of different types of content. FSIM performs better for content of high SI and edge entropy index, whereas MS-SSIM stands out for those blocks with lower content feature values. Also, the metrics that exhibit the highest correlation values overall are FSIM, MSSIM and PSNR-HVSM. Based on this small study, we will consider the SOA objective metrics as content features that will help to develop a multi-metric approach for quality prediction closer to the human visual experience.

Table 5
Distribution of blocks per dataset.

Cluster	Training	Testing	Total
K0	15.2%	21.8%	17%
K1	8.7%	19.1%	12%
K2	7.1%	9.4%	8%
K3	9.3%	.2%	7%
K4	44.7%	14.6%	36%
K5	9.3%	29.4%	15%
K6	5.7%	5.5%	6%

4. BVQA: Block-level multi-metric fusion for video quality assessment

4.1. Dataset partition for training and testing

Prior to the multi-metric fused model development and to avoid over-fitting, we divide our dataset into two parts: 70% of the data are used for fitting our models and 30% for testing purposes. In addition to this, the partitioning is performed on a sequence level ensuring that only blocks from sequences of the training set are extracted for training purposes (and the same for testing). Due to the small number of sequences, a selection method based on the content characteristics is suggested. Particularly, a scoring system is formulated to indicate which sequences can best represent the total population. This scoring is based on uniformity and coverage [30] of the block data in each sequence against the total population of data points available.

Table 4 lists the uniformity and coverage of the three chosen content attributes. The uniformity is calculated using the entropy of the histogram bins that evenly span the whole set of sequences, whereas the coverage is calculated for the normalized dimensions of the three content features, as explained in [30]. Finally, the product of the mean value of the uniformity for each dimension and of the coverage generates a score (i.e. Score $U_{mean} \cdot T$) that indicates how well each sequence represents the population. Out of this score we select the six sequences that are located within the 35th and 65th percentile. This decision derives from the motivation to partition the dataset in two representative sets suitable for training and testing. Choosing sequences from the same percentile for training (i.e. featuring great coverage and uniformity) would result in poor performance in testing.

In Table 4, the chosen sequences are highlighted in light grey. As can be observed, the choice of the middle six sequences based on the score allowed the division of the population into training and testing sets that adequately represent the whole population and are referred to in Table 4 as "Training Seqs" and "Testing Seqs". Indeed, while the overall population scores a total of .457, the training and testing subsets follow closely with a score .436 and .332 respectively. Next, we inspect the distribution of the blocks across the 7 clusters in Table 5. It can be seen that although the training set adequately represents the total population, the testing set includes a higher percentage of blocks in some clusters (e.g. K5) against others (e.g. K3). This is expected to impact on the performance of the prediction models between the testing and the training set.

4.2. Model fitting

Our hypothesis here is that a better performing quality assessment metric can result by combining several other state of the art metrics in a content-dependent manner. The fusion of the multiple metrics is achieved by applying multivariate fitting of the objective metrics and the content features. We have designed different families of predictors that use a different combination and number of inputs, that as a consequence also result in different computational complexity. The first two families, LL and LH, are a result of a linear combination of input metrics. Particularly, LL models are a result of the linear combination of up to three state-of-the-art quality metrics and LH models are a result of all considered quality metrics. The third family of models, NL, are non-linear combinations of quality metrics and content features. The software used for the model fitting purpose is Eureqa Pro software [31, 32]. We would like to note that we used a justified hold-out method instead of a random N-fold cross-validation (see Section 4.1).

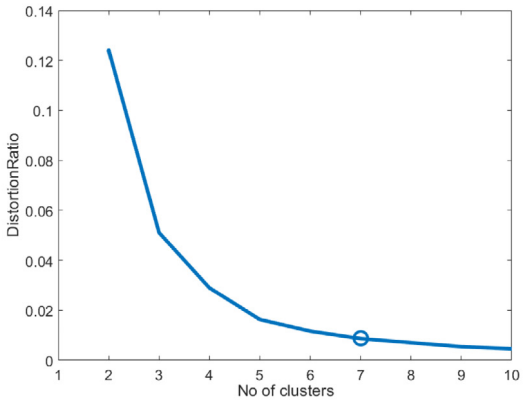
The predicted DMOS, $DMOS_p$, is continuous and limited within the range $[0, 5]$ according to the reported range for the collected DMOS values. The fitted models in all three families of predictors are reported in Table 6.

In order to assess the goodness of fit of the models, the following metrics are reported in Table 7: R^2, Lin, MSE and mean absolute error (MAE). We observe that the models fit reasonably well and provide good DMOS prediction. LL predictors consider only positive weight coefficients, resulting in solutions that are simple linear combinations of a few metrics. This limitation is removed for the LH family of predictors, where all metrics are considered for the first order linear fitting. This introduces a clear computational overhead as all metrics have to be calculated within the RDO. Finally, for the NL predictors non-linear formulas are examined that can potentially combine all

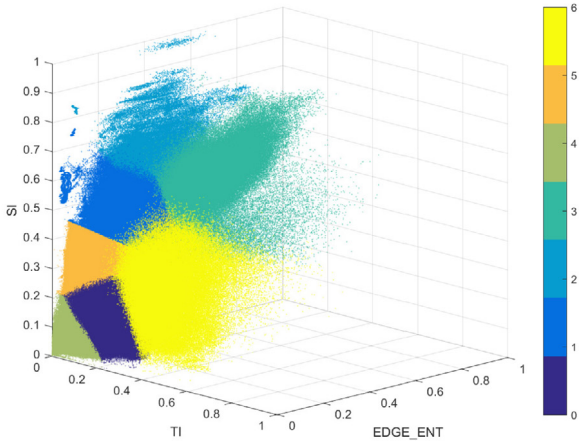
Table 6
BVQA Models.

	Cl.	$DMOS_p$ Model
BVQA-LL	K0	$5 - 1.025 \cdot M_1 - 2.072 \cdot M_2 - 6.057 \cdot M_3$
	K1	$5 - .541 \cdot M_2 - .592 \cdot M_4 - 4.359 \cdot M_5$
	K2	$5 - .786 \cdot M_4 - 2.096 \cdot M_7 - 2.558 \cdot M_5$
	K3	$5 - 1.714 \cdot M_7 - 3.265 \cdot M_5$
	K4	$5 - .811 \cdot M_1 - 1.769 \cdot M_2 - 6.757 \cdot M_3$
	K5	$5 - .823 \cdot M_5 - 1.389 \cdot M_2 - 2.119 \cdot M_1 - 2.677 \cdot M_3$
	K6	$5 - 2.766 \cdot M_3 - 3.581 \cdot M_1$
BVQA-LH	K0	$5 + 33.989 \cdot M_4 + 28.285 \cdot M_6 + .462 \cdot M_5 - 1.296 \cdot M_2 - 1.934 \cdot M_1 - 3.198 \cdot M_7 - 63.533 \cdot M_3$
	K1	$5 + 37.308 \cdot M_4 + 1.186 \cdot M_7 - .263 \cdot M_2 - .636 \cdot M_1 - 2.461 \cdot M_5 - 1.656 \cdot M_6 - 3.271 \cdot M_3$
	K2	$5 + 3.785 \cdot M_6 + 2.718 \cdot M_3 + .364 \cdot M_2 - .001 \cdot M_1 - 1.690 \cdot M_7 - 3.4 \cdot M_5 - 7.612 \cdot M_4$
	K3	$5 + 3.028 \cdot M_4 + 1.168 \cdot M_2 - .011 \cdot M_1 - .881 \cdot M_3 - 2.124 \cdot M_5 - 2.53 \cdot M_7 - 4.361 \cdot M_6$
	K4	$5 + 56.792 \cdot M_4 + 2.851 \cdot M_6 + 2.065 \cdot M_7 - .828 \cdot M_5 - .922 \cdot M_2 - 3.03 \cdot M_1 - 64.715 \cdot M_3$
	K5	$5 + 34.499 \cdot M_4 + 4.268 \cdot M_7 + .794 \cdot M_6 - .996 \cdot M_2 - 1.868 \cdot M_1 - 3.256 \cdot M_5 - 42.748 \cdot M_3$
	K6	$5 + 101.003 \cdot M_4 + 2.535 \cdot M_7 + .573 \cdot M_5 - .256 \cdot M_2 - 1.413 \cdot M_1 - 32.252 \cdot M_6 - 76.128 \cdot M_3$
BVQA-NL	K0	$.136/(.162 \cdot M_3 + 575.559 \cdot SI \cdot M_2 \cdot M_3^5 - EDGE\_ENT \cdot M_2)$
	K1	$.892/(74.811 \cdot M_3^2 - 71.726 \cdot M_4^2)$
	K2	$(1.03 + 1.771 \cdot SI \cdot TI \cdot M_7 - TI - M_5^2)/M_3$
	K3	$M_4/(.917 + 11.458 \cdot M_6 \cdot M_2 \cdot M_3^3 \cdot M_7^2 - M_7 - M_2)$
	K4	$(2.167 + 44.807 \cdot M_4 - TI - 47.642 \cdot M_3)/(M_6 + M_2 - \cdot EDGE\_ENT)$
	K5	$(M_7 - 3.372 \cdot M_1 \cdot M_3^2)/(.236 \cdot M_5 + M_5 \cdot M_2) - EDGE\_ENT$
	K6	$.874 + SI + 3.565 \cdot M_5 \cdot M_4 \cdot M_7^2 + -.008/(M_4^2 - 1.015 \cdot M_3^2) - SI \cdot M_2 - 5.083 \cdot M_6$

where M_1:MSSSIM, M_2:VIFP, M_3:PSNRHVSM, M_4:PSNRHVS, M_5:FSIM, M_6:PSNR, and M_7:SSIM .



(a) The optimal number of clusters equals to seven as observed by illustrating the elbow method on the feature clustering.



(b) Plot of the resulting clustered blocks in the three content feature dimensions.

Fig. 2. Results of the clustered blocks in the three content feature dimensions.

Table 7
Goodness of fit of BVQA Models and relative average complexity.

	BVQA-LL				BVQA-LH				BVQA-NL			
Cluster	R^2	Lin	MSE	MAE	R^2	Lin	MSE	MAE	R^2	Lin	MSE	MAE
K0	.749	.867	.169	.316	.774	.894	.153	.293	.817	.904	.123	.249
K1	.704	.839	.034	.135	.564	.867	.058	.196	.745	.864	.034	.124
K2	.833	.913	.017	.102	.804	.923	.020	.10	.860	.928	.015	.091
K3	.708	.841	.015	.084	.691	.851	.016	.077	.738	.859	.014	.077
K4	.533	.736	.279	.399	.470	.810	.316	.408	.630	.811	.221	.33
K5	.566	.754	.185	.304	.585	.826	.177	.316	.692	.833	.131	.238
K6	.506	.712	.077	.165	.468	.779	.083	.216	.646	.804	.055	.136
Rel. Cmplx	5.27				11.02				5.74			

seven metrics as well as the three primary content features. In this case, the goodness of fit metrics improve for the NL predictors. Although the NL predictors could result in an arithmetically more complex solution since they include floating point multiplication and division of the individual features and metrics, during the fitting only a subset of the features was selected resulting in an overall execution time that is lower than that of the LH models. To provide an indication of the computational complexity of the proposed models, we followed the same approach as earlier in Table 1. Thus, in the last row of Table 7 we are reporting the relative average complexity of the three models with reference to the minimum objective metric execution time, aka PSNR. As expected by considering the execution times of the different SOA quality metrics, BVQA-LL and BVQA-NL models are the fastest to compute due to the smaller number of input metrics.

4.3. Model validation

Fig. 3 illustrates the process of using the proposed block-level quality assessment to predict the expected perceived quality per block. After extracting the content features at a block level from the original video blocks, the blocks are classified in one of the seven clusters identified above. Then, the objective quality metric values are computed using the encoded video blocks. The content feature values, the assigned class and the quality metric values are fed into the BVQA models and the perceived video quality per block is estimated.

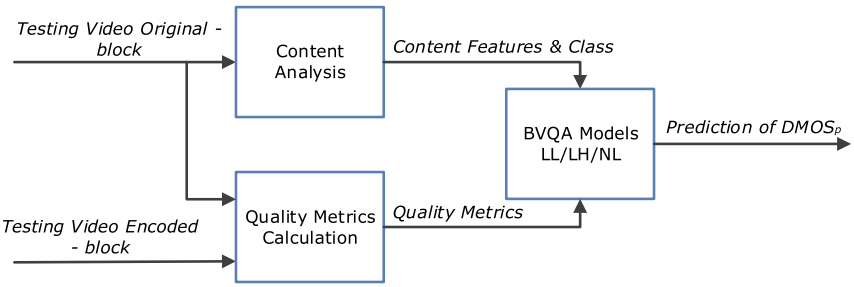


Fig. 3. Block diagram of the BVQA deployment.

Table 8
Linear and Rank correlation performance of SOA metrics and BVQA on the Training Set from BVI_Texture.

		FSIM	MSSSIM	PSNR-HVSM	BVQA-LL	BVQA-LH	BVQA-NL
Lin	Blk	.660	.732	.342	.775	.808	**.833**
	Frm	.764	.849	.443	.863	.894	**.912**
	Seq	.773	.862	.447	.875	.911	**.926**
Rnk	Block	.598	.675	.596	.728	.740	**.750**
	Frm	.706	**.850**	.635	.830	.837	.845
	Seq	.725	**.870**	.644	.844	.851	.856

Table 9
Linear and Rank correlation performance of SOA metrics and BVQA on the Testing Set of BVI_Texture.

		FSIM	MSSSIM	PSNR-HVSM	BVQA-LL	BVQA-LH	BVQA-NL
Lin	Blk	.568	.459	**.649**	.537	.494	.506
	Frm	.659	.535	**.718**	.590	.542	.568
	Seq	.672	.554	**.726**	.601	.558	.590
Rnk	Block	.632	.630	.741	.692	.722	**.789**
	Frm	.692	.727	.803	.743	.787	**.825**
	Seq	.686	.720	.793	.731	.783	**.804**

For the BVQA model validation, we evaluate the performance of the method against the best performing metrics from Table 3, namely FSIM, MS-SSIM, and PSNR-HVSM. In Table 8, we list the linear and rank correlation coefficients between the original [24] and the predicted $DMOS_p$ using BVQA models for both the training set. For the training set, we identify that the model has been fitted correctly for each cluster by observing the first couple of columns. Each fitting solution (LL, LH and NL models) shows a linear and rank correlation between .78-.83 and .73-.75 at the block level, respectively. As can be observed, the correlation values increase at the frame (.86-.91 and .83-.85) and at the sequence level (.88-.93 and .84-.86). This shows the effectiveness of the method as high correlation with the DMOS scores is achieved at the sequence level overall.

To further verify the BVQA models, we use two other datasets. The first is the testing set of BVI_Texture and the other is the VQEG-HD3 dataset. We have selected this dataset as it complies with the assumption we made for sequences without an apparent viewing task and it is annotated with subjective scores. It is expected that the model performance will deviate for these two datasets compared to the training set mainly because the available number of sequences annotated with subjective scores is not high and diverse enough to cover the feature and objective quality metrics space.

The results of deploying the BVQA models for the testing sequences of BVI_Texture are reported in Table 9. As anticipated, the correlation values drop in the testing set. However, BVQA outperforms the state of the art quality metrics in terms of rank correlation. The drop of performance in the testing set is a natural effect of the variability and randomness of the selected blocks from the video sequences, as well as of the small number of sequences available for the training.

Table 10
Linear and Rank correlation performance of SOA metrics and BVQA on VQEG-HD3.

		FSIM	MSSSIM	PSNR-HVSM	BVQA-LL	BVQA-LH	BVQA-NL
Lin	Blk	.413	.378	.462	**.485**	.477	.454
	Frm	.562	.455	.523	**.563**	.548	.537
	Seq	.742	.768	.658	.774	**.791**	.788
Rnk	Block	.470	.468	.477	.496	**.500**	.499
	Frm	**.595**	.587	.551	.583	.575	.574
	Seq	.769	**.867**	.716	.786	.832	.816

Finally, we present the results on another dataset from Video Quality Expert Group (VQEG) with HD videos, the VQEG-HD3 dataset [33]. For this dataset, as reported in Table 10, all tested metrics achieve lower linear and rank correlation values compared to those from the testing sequences in BVI_Texture dataset (see Table 9). This is expected due to the different content characteristics of this dataset. Nevertheless, in most cases, BVQA outperforms the state of the art objective quality metrics in this dataset.

5. Conclusion

We presented a multi-metric fusion approach, which delivers a video quality assessment method at a block level that correlates better with perceptual quality compared to the state-of-the-art objective metrics. This approach is a step towards combining several well-performing metrics into one, exploiting the advantages of using objectives metrics that are embeddable in the RDO process in a content-dependent manner. At the same time, the advantage of developing a block-level quality metric is that of using it within the RDO environment. The first results of BVQA are promising in terms of the correlation of the developed locally-acting quality metric with the overall perceived quality of the video. This allows us to argue that, within this group of content, this combination of metrics produces a quality estimate closer to the average experience. Consequently, the RDO is expected to be more efficient as it will be using a model that is more affected by a higher level of content awareness (what is around it) and not just by the content of the block.

6. Limitations and challenges for future work

Recently, with the aim to optimize the trade-off of the encoding pipeline and the streaming performance, the videos are split in "chunks" (often at a shot level) of a few seconds as proposed for example in [19,20]. The presented multi-metric fusion method is built on the assumption that for short videos that could represent one shot, we have homogeneity in terms of the scene content across all tested frames. We have also assumed for this work no apparent viewing patterns. It is however important to take into account the perceptual significance of specific parts of a frame either because of visual salience or/and the semantic importance. Thus, the challenge is to extend our method to take into account the perceptual importance of specific areas that might be points of interest for most viewers.

Furthermore, the results presented in this paper were based only on a limited number of sequences coming from two datasets in order to conform with the method assumptions. Then, we followed a hold-out validation method using a justified splitting of the sequences that was based on the relative coverage and uniformity of the low-level features of the dataset at a sequence level. The challenge arising form this is to further test the method against new datasets and perform a cross-validation with randomized splits.

Finally, the biggest challenge once BVQA is the natural step of integrating the proposed method in the RDO of a video encoder, and computing the effectiveness (gains both in quality and bit rate) and the efficiency (complexity overhead) of the BVQA-based optimized encodings.

Acknowledgements

The work was supported by the "Marie Skłodowska-Curie Actions - the EU Framework Programme for Research" ITN PROVISION, by the Engineering and Physical Sciences Research Council (EPSRC), UK EP/M000885/1 and by the Leverhulme Trust, UK.

References

[1] Z. Wang, A.C. Bovik, H.R. Sheikh, E.P. Simoncelli, Image quality assessment: from error visibility to structural similarity, Trans. Image Process. 13 (4) (2004) 600–612.
[2] D.R. Bull, Communicating Pictures, Academic Press, 2014.
[3] N. Ponomarenko, F. Silvestri, K. Egiazarian, M. Carli, J. Astola, V. Lukin, On between-coefficient contrast masking of DCT basis functions, in: Proc. of the 3rd Intern. Workshop on Video Processing and Quality Metrics, vol. 4, 2007.
[4] Z. Wang, E.P. Simoncelli, A.C. Bovik, Multiscale structural similarity for image quality assessment, in: 37th Asilomar Conf. on Signals, Systems and Computers, vol. 2, IEEE, 2004, pp. 1398–1402.
[5] H.R. Sheikh, A.C. Bovik, A visual information fidelity approach to video quality assessment, in: 1st Intern. Workshop on Video Processing and Quality Metrics for Consumer Electronics, 2005, pp. 23–25.
[6] L. Zhang, L. Zhang, X. Mou, D. Zhang, FSIM: A feature similarity index for image quality assessment, IEEE Trans. Image Process. 20 (8) (2011) 2378–2386.
[7] F. Zhang, D.R. Bull, A perception-based hybrid model for video quality assessment, Trans. Circuits Syst. Video Technol. 26 (6) (2016) 1017–1028.
[8] K. Seshadrinathan, A.C. Bovik, Motion-based perceptual quality assessment of video, in: IS&T/SPIE Electronic Imaging, Intern. Society for Optics and Photonics, 2009, p. 72400X.
[9] M.H. Pinson, S. Wolf, A new standardized method for objectively measuring video quality, IEEE Trans. Broadcast. 50 (3) (2004) 312–322.
[10] M. Vranješ, S. Rimac-Drlje, K. Grgić, Review of objective video quality metrics and performance comparison using different databases, Signal Process., Image Commun. 28 (1) (2013).
[11] J. Qi, X. Li, F. Su, Q. Tu, A. Men, Efficient rate-distortion optimization for HEVC using SSIM and motion homogeneity, in: 2013 Picture Coding Symposium, PCS.
[12] T. Zhao, K. Zeng, A. Rehman, Z. Wang, On the use of SSIM in HEVC, in: Signals, Systems and Computers, 2013 Asilomar Conf. on, IEEE, 2013, pp. 1107–1111.
[13] K. Naser, V. Ricordel, P. Le Callet, Experimenting texture similarity metric STSIM for intra prediction mode selection and block partitioning in HEVC, in: IEEE 19th Intern. Conf. on Digital Signal Processing, DSP, 2014, pp. 882–887.
[14] C. Yeo, H.L. Tan, Y.H. Tan, On rate distortion optimization using SSIM, IEEE Trans. Circuits Syst. Video Technol. 23 (7) (2013) 1170–1181.
[15] C. Yeo, H.L. Tan, Y.H. Tan, SSIM-based adaptive quantization in HEVC, in: IEEE Intern. Conf. on Acoustics, Speech and Signal Processing, ICASSP, IEEE, 2013, pp. 1690–1694.
[16] J. Korhonen, J. You, Improving objective video quality assessment with content analysis, in: Proc. of the 5th Intern. Workshop on Video Processing and Quality Metrics for Consumer Electronics, VPQM, Scottsdale, USA, 2010.
[17] Z. Li, A. Aaron, I. Katsavounidis, A. Moorthy, M. Manohara, Toward a practical perceptual video quality metric, 2016, https://medium.com/netflix-techblog/toward-a-practical-perceptual-video-quality-metric-653f208b9652, (Accessed 20 June 2017).
[18] S. Li, F. Zhang, L. Ma, K.N. Ngan, Image quality assessment by separately evaluating detail losses and additive impairments, IEEE Trans. Multimedia 13 (5) (2011) 935–949.
[19] S. Lederer, C. Müller, C. Timmerer, Dynamic adaptive streaming over HTTP dataset, in: Proceedings of the 3rd Multimedia Systems Conference, MMSys, ACM, 2012, pp. 89–94.
[20] J. De Cock, Z. Li, M. Manohara, A. Aaron, Complexity-based consistent-quality encoding in the cloud, in: IEEE International Conference on Image Processing, ICIP, 2016, pp. 1484–1488.
[21] I. Katsavounidis, Dynamic optimizer - a perceptual video encoding optimization framework, https://medium.com/netflix-techblog/dynamic-optimizer-a-perceptual-video-encoding-optimization-framework-e19f1e3a277f.
[22] A. Zabrovskiy, C. Feldmann, C. Timmerer, A practical evaluation of video codecs for large-scale HTTP adaptive streaming services, in: 25th IEEE International Conference on Image Processing, ICIP, 2018, pp. 998–1002.
[23] F. Mercer Moss, K. Wang, F. Zhang, R. Baddeley, D.R. Bull, On the optimal presentation duration for subjective video quality assessment, IEEE Trans. Circuits Syst. Video Technol. 26 (11) (2016) 1977–1987.
[24] M.A. Papadopoulos, F. Zhang, D. Agrafiotis, D. Bull, A video texture database for perceptual compression and quality assessment, in: Intern. Conf. on Image Processing, ICIP, IEEE, 2015, pp. 2781–2785.
[25] I. Recommendation, P.910, Subjective Video Quality Assessment Methods for Multimedia Applications, vol. 910, International Telecommunication Union, Geneva, Switzerland, 1999.
[26] G.N. Srinivasan, G. Shobha, Statistical texture analysis, in: Proc. of World Academy of Science, Engineering and Technology, vol. 36, 2008, pp. 1264–1269.
[27] D. Agrafiotis, D.R. Bull, N. Canagarajah, Enhanced spatial error concealment with directional entropy based interpolation switching, in: IEEE Intern. Symposium on Circuits and Systems, ISCAS, 2006.
[28] T.K. Moon, The expectation-maximization algorithm, IEEE Signal Process. Mag. 13 (6) (1996) 47–60.
[29] T.M. Kodinariya, P.R. Makwana, Review on determining number of cluster in k-means clustering, Intern. J. 1 (6) (2013) 90–95.
[30] S. Winkler, Analysis of public image and video databases for quality assessment, J. Sel. Topics Signal Process. 6 (6) (2012) 616–625.
[31] M. Schmidt, H. Lipson, Distilling free-form natural laws from experimental data, 324 (5923) (2009) 81–85.
[32] M. Schmidt, H. Lipson, Eureqa (Version 0.98 beta)[software], Nutonian, Somerville, Mass, USA, 2014.
[33] The Consumer Digital Video Library, http://www.cdvl.org/.