A Novel Methodology for Mapping Objective Video Quality Metrics to the Subjective MOS Scale

Arghir-Nicolae Moldovan, Ioana Ghergulescu, and Cristina Hava Muntean
School of Computing
National College of Ireland
Mayor Street, IFSC, Dublin 1, Ireland
e-mail: amoldovan@student.ncirl.ie; ioana.ghergulescu@ncirl.ie; cristina.muntean@ncirl.ie

***Abstract*—With the rapid growth in video-based services, and as users are becoming increasingly quality-aware, the reliable estimation of video quality has become extremely important. While a multitude of objective Video Quality Assessment (VQA) metrics with various performance and complexity have been proposed, the nonlinearity of video quality and the lack of clear interpretations of the metrics make difficult to understand how the objective metric values reflect the video quality as perceived subjectively in terms of Mean Opinion Scores (MOS). This paper proposes and evaluates a methodology for mapping objective VQA metric values to subjective MOS scores based on publicly available VQA databases. Three different databases were used for comparing the performance of various objective metrics and evaluating the proposed methodology.**

***Index Terms*—Video quality assessment (VQA), mean opinion score (MOS), subjective methods, objective metrics, video quality mapping, performance evaluation, public video quality databases.**

I. Introduction

OVER the past few years there has been an increasing adoption of in Internet multimedia services and applications such as video-on-demand, IPTV and video conferencing. It is expected that these services will continue to grow at a fast pace, with the mobile video in particular being estimated to increase 14-fold between 2013 and 2018, and to reach 69% of the global mobile Internet traffic by 2018 [1]. At the same time, mobile users are increasingly becoming quality-aware with the proliferation of high-definition multimedia content and high resolution device displays [2].

In this context, it has become extremely important to be able to reliably estimate the quality of multimedia services, in order to provide the users with a good Quality of Experience [3]. Researchers and industry alike have recognised this, and much work has been conducted in order to develop objective VQA metrics that can be implemented with multimedia services for automatic estimation of video quality without the need for the users to provide their opinion [4].

However the usability of objective VQA metrics is limited by the lack of clear interpretations of their values and how these reflect the subjective user-perceived video quality expressed on the MOS scale (i.e., 1 – bad, 2 – poor, 3 – fair, 4 – good, and 5 – excellent) [5]. This paper addresses the issue by proposing a methodology for mapping the values of objective VQA metrics to subjective MOS scores. The methodology builds on previous research works in the area, and makes use of data from public VQA databases [6]. The quality estimation performance of six different full-reference objective VQA metrics was evaluated on three different databases, and the proposed methodology was exemplified and evaluated for the best performing metric.

The rest of the paper is structured as follows. Section II briefly presents the state of the art in the area of video quality assessment. Section III presents the proposed methodology for mapping objective VQA metric values to subjective MOS scores, while Section IV presents the evaluation results. Section V draws conclusions and presents future work directions.

II. Video Quality Assessment Background

There are two main approaches for assessing the video quality, namely subjective VQA methods and objective VQA metrics [7].

A. Subjective VQA Methods

Subjective methods for video quality assessment are considered the most accurate and reliable way for assessing the video quality. Several subjective VQA methods are standardised by ITU in the recommendations ITU-R Rec. BT.500 [8] for television and ITU-T Rec. P.910 [9] for multimedia applications. These standards provide useful guidelines and instructions regarding the selection of the subjects and of the test material, the setup of the test environment, the rating scales to be used for assessment, as well as the methods for analysing the data.

There are two major subjective VQA approaches based on the presentation of test sequences: *double stimulus (DS)* and *single stimulus (SS)*. In case of DS methods such as the Degradation Category Rating (DCR) [9], viewers are presented with pairs of sequences and are asked to rate either each sequence individually or to rate the difference between them. SS methods such as the Absolute Category Rating (ACR) [9],

Manuscript received 27 April 2014. This work was supported by Irish Research Council's Embark Postgraduate Scholarship Scheme.

enable a higher number of test sequences to be rated in the same testing duration.

Various discrete (e.g., 1 – bad to 5 – excellent) or continuous (e.g., 1 to 100) scales can be used for rating purposes, although when compared against each other they were shown to lead to very similar results as long as a careful test design is conducted and clear information is provided to participants [10].

A multitude of subjective studies were conducted by various research groups such as for example the Video Quality Experts Group (VQEG)[1] with some of these being overviewed in [11]. Moreover, the results of many studies have been made accessible to other researchers as public VQA databases [6].

B. Objective VQA Metrics

As the need for subjects opinion renders unfeasible the use of subjective VQA methods in real-world applications, there has been much research work on proposing objective VQA metrics that quantify mathematically the video quality. Objective VQA metrics are wildly employed for assessing the video quality in both prototype-based [12] and simulation-based solutions [13]. Relative to the presence of the original reference video stream unaffected by the factors under test, the existing objective VQA metrics can be classified in: *no-reference (NR)*, *reduced-reference (RR)* and *full-reference (FR)* metrics [7].

NR objective VQA metrics have a high flexibility as they do not require the presence of the reference video. These metrics quantify the video quality based on various factors such as blockiness (i.e., distortion common to block-based compression algorithms such as H.264) blurring, jerkiness, ringing, etc. [14].

RR objective VQA metrics aim to provide a compromise between the measurement accuracy and flexibility of use. They use only some information extracted from the reference video, such as the amount of motion or spatial detail, which have lower bitrate and are more feasible to be transmitted over the communication channel [15].

FR objective VQA metrics enable the highest quality estimation performance, but they require the presence of the reference video, as well as precise spatial and temporal synchronisations, and luminance and colour calibration between the original and the impaired videos. The Peak Signal-to-Noise Ratio (PSNR), is the most widely used FR objective metric due to its simplicity, even though is often criticised for having a poor correlation with the subjective tests [16]. A multitude of more complex metrics based on natural visual characteristics, or that aim to model the Human Visual System (HVS), have also been proposed [4]. These metrics incorporate factors such as colour perception, contrast sensitivity or pattern masking, with one example being the Structural Similarity Index (SSIM) metric and its different variations [17].

Table I
EXAMPLES OF SOLUTIONS FOR MAPPING THE VALUES OF PSNR AND SSIM OBJECTIVE VQA METRICS TO SUBJECTIVE MOS SCORES, THAT WERE FOUND IN THE LITERATURE REVIEW.

MOS	**PSNR** [19]	**PSNR** [20]	**SSIM** [20]
5 (Excellent)	≥37	≥45.0	≥0.99
4 (Good)	≥31 & <37	≥33.0 & <45.0	≥0.95 & <0.99
3 (Fair)	≥25 & <31	≥27.4 & <33.0	≥0.88 & <0.95
2 (Poor)	≥20 & <25	≥18.7 & <27.4	≥0.50 & <0.88
1 (Bad)	<20	<18.7	<0.50

III. Methodology for Mapping Objective VQA Metric Values to the Subjective MOS Scale

A. Overview

The subjective user-perceived video quality is usually expressed using discrete MOS scales, such as the 5-point scale (i.e., 1-Bad, 2-Poor, 3-Fair, 4-Good and 5-Excellent). As opposed, the objective VQA metrics return quality values on various specific continuous scales such as for example 1-100 in case of the PSNR metric [16], or 0-1 in case of the SSIM metric [17].

Since the video quality is usually a non-linear measure [18], it is difficult to interpret the objective quality values in terms of MOS scores. Moreover while there is much research work being made on proposing objective quality metrics, there is little research that provides clear mappings of the objective VQA metric values to subjective user-perceived quality levels expressed on the MOS scale. This is because the evaluation of objective metrics is usually limited only to indicating how well they correlate with the subjective MOS scores.

Table I presents two mapping solutions that were found in the literature for the PSNR metric [19], [20], as well as a mapping for the SSIM metric [20].

The steps of the proposed methodology for mapping the values of an objective VQA metric to the subjective MOS scores based on subjective data from public VQA databases, are presented next.

B. Mapping Methodology Steps

The proposed methodology for mapping objective VQA metric values to subjective MOS scores consists of the following steps:

Step 1: Identify and select suitable VQA datasets by consulting available public VQA databases, or alternatively obtain the data through a comprehensive subjective study. The datasets should provide a clear description on how the data was collected and processed, should consider multiple sequences covering a broad range of content characteristics (e.g., news, sports, cartoons, etc.), should provide both the reference sequences and the test sequences affected by the various artefacts tested, and should provide at least the subjective MOS scores and standard deviations if not all the individual participants' ratings.

[1] Video Quality Experts Group (VQEG), http://www.its.bldrdoc.gov/vqeg/vqeg-home.aspx.

Step 2: Convert all subjective ratings to the same scale, if necessary when using data from multiple VQA datasets. The set of subjective ratings corresponding to the test sequences form a VQA dataset s, can be expressed conceptually as in (1).

$$MOS_s = \left\{ MOS_{sk}, k = \overline{1, N_s} \right\} \tag{1}$$

where k represents the test sequence index, while N_s represents the number of test sequences from the dataset s.

Step 3: Compute the values for the objective VQA metric for which the mapping is desired, for all the test sequences. Similarily, the set of objective VQA values (OM) corresponding to the test sequences form a VQA dataset s, can be expressed conceptually as in (2).

$$OM_s = \left\{ OM_{sk}, k = \overline{1, N_s} \right\} \tag{2}$$

where k represents the test sequence index, while N_s represents the number of test sequences from the dataset s.

Step 4: Perform nonlinear regression between the subjective MOS scores and objective VQA values. This step is necessary as subjective rating data are often compressed at the ends of the rating scales, and it is not reasonable for objective models of video quality to mimic this weakness of subjective data [21]. A nonlinear mapping function that is commonly used as it was found to perform well empirically [21], [22] is the cubic polynomial, expressed as in (3).

$$MOS_s^p = a \cdot OM_s^3 + b \cdot OM_s^2 + c \cdot OM_s + d \tag{3}$$

where MOS_s^p represents the predicted quality by the particular objective VQA metric, while the a, b, c and d constants are obtained by fitting the function to the data $[MOS_s, OM_s]$.

Step 5: Evaluate the mapping performance by analysing the goodness of the nonlinear fitting indicated by the R^2 measure. An R^2 value closer to 1 indicates a better performance.

Step 6: Compute lower and upper objective VQA metric threshold values corresponding to the different MOS quality levels, by applying inverse interpolation on the nonlinear regression model.

IV. Evaluation of the Mapping Methodology

A. VQA Databases

Subjective data from three public video quality assessment databases were used in order to evaluate the quality estimation accuracy of six different objective VQA metrics, as well as the proposed mapping methodology for the best performing metric. The three databases were selected because they considered the impact of distortions that are introduced by current generation video codecs such as H.264/MPEG-4 AVC. Moreover, each database considered multiple test sequences with different content characteristics, and combined the databases cover a broad range of video resolutions.

The ETFOS CIF Video Quality (ECVQ) database [23] contains 8 progressive reference sequences in the raw YUV 4:2:0 format, each of them having a CIF (352×288 pixels) resolution, a 25 fps framerate, and a 12 seconds duration. Moreover, the database contains 90 test sequences, compressed at different bitrates from 73 kbps to 827 kbps, approximately half of them using the H.264/MPEG-4 AVC codec and the other half using the MPEG-4 Visual codec.

The ETFOS VGA Video Quality (EVVQ) database [23] contains 8 progressive reference sequences in the raw YUV 4:2:0 format, each of them having a VGA (640×480 pixels) resolution, a 25 fps framerate, and a 12 seconds duration. Similarly to the ECVQ database, the EVVQ database contains 90 test sequences compressed using the H.264 and the MPEG-4 Visual codecs, but at bitrates between 261 kbps to 1737 kbps.

The LIVE Mobile Video Quality Assessment database [24], consists of 10 progressive reference sequences in the raw YUV 4:2:0 format, each of them having a HD 720p (1280×720 pixels) resolution, a 30 fps framerate, and a 15 seconds duration. The video quality subjective data corresponding to 40 test sequences compressed using the H.264 codec at bitrates between 0.7 Mbps to 6 Mbps, was used for evaluating the performance of various objective metrics. The 40 sequences were rated on a smaller screen smartphone with a resolution of 960×540, while half of them were also rated on a larger screen tablet with a resolution of 1280×800.

Single stimulus with hidden reference subjective evaluation methods were used for all three databases. The databases contain the MOS scores for each test sequence averaged across all participants as in (4).

$$MOS_k = \frac{1}{N}\sum_{i=1}^{N} d_{ik} \tag{4}$$

where k is the test sequence index, i is the participant index, N is the number of participants, while d_{ik} are difference scores obtained on a per subject per test sequence basis.

The difference scores for the ECVQ and EVVQ databases were computed as in (5) and could take values between 0 to 100, with 100 being the highest quality [23]. As opposed, the difference scores for the LIVE Mobile database were computed as in (6) and could range between 0 and 5, with 0 being the highest quality [24].

$$d_{ik} = s_{ik} - s_{ik}(ref) + 100 \tag{5}$$

$$d_{ik} = s_{ik} - s_{ik}(ref) \tag{6}$$

where s_{ik} represents the rating of the participant i for the k test sequence, while $s_{ik}(ref)$ represents the participant's rating for the reference video from which the sequence was obtained.

To have all data on the same scale, the MOS scores for the LIVE Mobile database were converted to the same 0-100 scale as for the ECVQ and EVVQ databases. This was done according to the formula from (7).

$$MOS_k = (5 - MOS_k) \cdot 20 \tag{7}$$

Table II
Objective VQA metrics's quality estimation accuracy results for the ECVQ, EVVQ and LIVE Mobile datasets, as indicated by the Pearson Linear Correlation Coefficient (PLCC).

	ECVQ			**EVVQ**			**LIVE Mobile (H.264)**			
VQA Metric	**H.264**	**MPEG-4 Visual**	**All**	**H.264**	**MPEG-4 Visual**	**All**	**Mobile**	**Tablet**	**All**	**All Data**
PSNR	0.7313	0.7233	0.7377	0.677	0.7629	0.7271	0.8392	0.8456	0.8164	0.6807
PSNR-HVS	0.8058	0.7849	0.8006	0.7514	0.8421	0.8064	0.8519	0.8453	0.8252	0.7838
PSNR-HVS-M	0.8909	0.7724	0.8301	0.8237	0.9108	0.8730	0.8927	0.9202	0.8785	0.8592
SSIM	0.9070	**0.8868**	**0.8987**	0.8217	0.8904	0.8597	0.7698	0.6203	0.7064	0.7754
MS-SSIM	**0.9269**	0.8618	0.8960	**0.8938**	**0.9151**	**0.9049**	0.8734	0.8509	0.8512	0.8850
VIFp	0.8851	0.8706	0.8806	0.8486	0.8793	0.8688	**0.9209**	**0.9827**	**0.9358**	**0.8986**

Table III
Objective VQA metrics's quality estimation accuracy results for the ECVQ, EVVQ and LIVE Mobile datasets, as indicated by the Root Mean Square Error (RMSE).

	ECVQ			**EVVQ**			**LIVE Mobile (H.264)**			
VQA Metric	**H.264**	**MPEG-4 Visual**	**All**	**H.264**	**MPEG-4 Visual**	**All**	**Mobile**	**Tablet**	**All**	**All Data**
PSNR	11.9184	11.095	11.5308	11.225	9.3459	10.4085	12.3885	11.8356	13.072	14.0054
PSNR-HVS	10.3469	9.9559	10.2344	10.0633	7.7968	8.967	11.9288	11.8469	12.7864	11.8726
PSNR-HVS-M	7.9362	10.2049	9.5232	8.6481	5.967	7.3938	10.2654	8.6784	10.8159	9.7819
SSIM	7.3602	**7.4242**	**7.4905**	8.692	6.5786	7.7447	14.5386	17.3899	16.0237	12.0724
MS-SSIM	**6.56**	8.1497	7.5841	**6.8393**	**5.8282**	**6.4519**	11.0936	11.6467	11.879	8.9026
VIFp	8.1332	7.9048	8.0927	8.0689	6.8839	7.5084	**8.8782**	**4.109**	**7.9806**	**8.3878**

B. Objective Metrics Computation

The Video Quality Measurement Tool (VQMT) [25] version 1.1, provided by the Multimedia Signal Processing Group (MMSPG) at Ecole Polytechnique Fédérale de Lausanne (EPFL) was used in order to compute the following six full-reference objective VQA metrics: Peak Signal-to-Noise Ratio (PNSR) [16], Structural Similarity Index (SSIM), Multi-scale Structural Similarity Index (MS-SSIM) [26], Visual Information Fidelity pixel domain version (VIFp) [27], PSNR taking into account Contrast Sensitivity Function (CSF) (PSNR-HVS) [28], and PSNR taking into account Contrast Sensitivity Function (CSF) and between-coefficient contrast masking of DCT basis functions (PSNR-HVS-M) [29].

Each objective VQA metric is computed for all the test sequences from the three databases. For each test sequence the metrics are computed relative to the source uncompressed clip. This is done on a frame-by-frame basis, while the per-sequence value of an objective VQA metric is taken as the average across all video frames. The MMSP VQMT tool has the advantage of being command line and thus it can be used for batch computation in case of multiple test sequences.

C. Objective Metrics Performance Comparison Results

The video quality prediction performance of the considered objective VQA metrics was compared in terms of accuracy, monotonicity, and consistency [4], [22]. The prediction accuracy is usually quantified through the Pearson Linear Correlation Coefficient (PLCC) and the Root Mean Square Error (RMSE). The Spearman's Rank Ordered Correlation Coefficient (SROCC) is usually used in order to quantify the monotonicity of the objective metric predictions with respect to human scores. The Outlier Ratio (OR) defined as the ratio of the number of predictions outside the range of ±2 times the standard deviations of the subjective results, is usually used as a measure of the prediction consistency (i.e., the degree to which the metric maintains the prediction accuracy).

The four metrics were computed after performing a nonlinear regression using a cubic polynomial as the one expressed in (3). The non-linear regression and the performance measures computation was performed individually for each database (i.e., ECVQ, EVVQ, and LIVE Mobile), for individual subsets of each database (i.e., H.264 vs. MPEG-4 Visual for the ECVQ and EVVQ databases, Mobile vs. Tablet for the LIVE Mobile database), as well across all the combined data of the three databases. The data processing and the statistical analysis were conducted using the R v.3.01 statistical software[2].

The performance comparison results of the six objective VQA metrics in terms of the PLCC, SROCC RMSE and OR measures are presented in Tables II to V respectively. The results show that overall across the combined data of the three datasets the VIFp metric offers the best performance in terms of prediction accuracy (highest PLCC and lowest RMSE),

[2] The R Project for Statistical Computing, http://www.r-project.org/.

Table IV
OBJECTIVE VQA METRICS'S QUALITY ESTIMATION MONOTONICITY RESULTS FOR THE ECVQ, EVVQ AND LIVE MOBILE DATASETS, AS INDICATED BY SPEARMAN'S RANK ORDERED CORRELATION COEFFICIENT (SROCC).

	ECVQ			**EVVQ**			**LIVE Mobile (H.264)**			
VQA Metric	**H.264**	**MPEG-4 Visual**	**All**	**H.264**	**MPEG-4 Visual**	**All**	**Mobile**	**Tablet**	**All**	**All Data**
PSNR	0.7815	0.7265	0.7617	0.7112	0.7758	0.7635	0.8270	0.8150	0.8051	0.7306
PSNR-HVS	0.7973	0.7332	0.7769	0.7712	0.8510	0.8273	0.8290	0.8150	0.8054	0.8018
PSNR-HVS-M	0.8439	0.6977	0.7976	0.8295	0.9154	0.8793	0.8926	0.9038	0.8835	0.8533
SSIM	**0.8961**	**0.9066**	**0.9161**	0.7827	0.8968	0.8647	0.6857	0.4060	0.6281	0.7753
MS-SSIM	0.8875	0.8157	0.8771	**0.8757**	**0.9306**	**0.9112**	0.8756	0.8511	0.8644	0.8852
VIFp	0.8766	0.8493	0.8745	0.8234	0.8995	0.8751	**0.9154**	**0.9744**	**0.9311**	**0.8877**

Table V
OBJECTIVE VQA METRICS'S QUALITY ESTIMATION MONOTONICITY RESULTS FOR THE ECVQ, EVVQ AND LIVE MOBILE DATASETS, AS INDICATED BY THE OUTLIER RATIO (OR).

	ECVQ			**EVVQ**			**LIVE Mobile (H.264)**			
VQA Metric	**H.264**	**MPEG-4 Visual**	**All**	**H.264**	**MPEG-4 Visual**	**All**	**Mobile**	**Tablet**	**All**	**All Data**
PSNR	0.0465	0.0000	0.0222	0.0000	0.0000	0.0000	0.0000	0.0000	0.0333	0.0250
PSNR-HVS	0.0000	0.0213	0.0111	0.0000	0.0000	0.0000	0.0000	0.0000	0.0333	0.0083
PSNR-HVS-M	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0042
SSIM	0.0000	0.0213	0.0111	0.0000	0.0000	0.0000	0.0500	0.1000	0.0500	0.0125
MS-SSIM	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0333	0.0042
VIFp	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	**0.0000**

prediction monotonicity (highest SROCC) and prediction consistency (lowest OR). On the overall data the MS-SSIM metric performs only slightly worse, while the PSNR metric had the lowest performance among the considered metrics.

Looking individually across the three different databases the results show that for the ECVQ database, containing sequences with the lowest resolution 352×288 pixels, the best performing metric is SSIM. For the EVVQ database that contains sequences with the middle resolution 640×480 pixels, the best performing metric is MS-SSIM, whereas the LIVE Mobile database that contains sequences with the highest resolution 1280×720 pixels, the best performing metric is VIFp.

D. Mapping the VIFp Metric to the Subjective MOS Scale

The best performing objective VQA metric VIFp, was chosen in order to illustrate how its values can be mapped to the subjective MOS scores, when the proposed mapping methodology is used.

Following the mapping methodology steps presented in section III-B, first the subjective data for the LIVE Mobile database was converted according to the formula from (7), to the same $0-100$ scale used by the ECVQ and EVVQ databases (*Step 2*). Moreover, the values of the VIFp metric that are expressed on a continuous scale between 0-1 were computed (*Step 3*). After that, a nonlinear regression using a cubic polynomial function as described by the formula from (3), was performed between the subjective MOS scores and the values of the VIFp objective metric (*Step 2*).

Figure 1 presents the nonlinear fitted cubic polynomial function used by the mapping methodology in order to map the values of the VIFp full-reference objective VQA metric to the MOS scale. The combined data from all three VQA databases was used for the mapping. The figure also presents the 95% confidence bands, as well as the four coefficients of the nonlinear model. The *nls* function from the R statistical software was used for computing the optimum coefficients that minimise the least square errors between the vector of objective VIFp values and the vector of subjective MOS values. Analysing the goodness of fit for the nonlinear regression model (*Step 5*), the results show that the mapping performance is as high as 80.76% ($R^2 = 0.8075, p <$ 2.2e-16, $F =$ 330, $DF = 236$).

Table VI presents the mapping of the VIFp objective VQA metric values to the subjective MOS scores, obtained by applying inverse interpolation on the nonlinear model (*Step 6*). Following the recommendations on video quality assessment scales from ITU [8], the continuous $0-100$ MOS scale was converted to the discrete $1-$Bad to $5-$Excellent MOS scale, by dividing it into 5 equal intervals of length 20 and assigning a discrete quality level to each interval. As illustrated in Figure 1 the lower VIFp threshold values for the 2, 3, 4 and 5 user-perceived MOS quality levels, correspond to the 20, 40, 60 and 80 levels on the continuous scale.

Table VII
VIFP MAPPING'S QUALITY ESTIMATION ACCURACY (PLCC AND RMSE), MONOTONICITY (SROCC) AND CONSISTENCY (OR) RESULTS FOR THE ECVQ, EVVQ AND LIVE MOBILE DATASETS.

	ECVQ			EVVQ			LIVE Mobile (H.264)			
Performance Measure	**H.264**	**MPEG-4 Visual**	**All**	**H.264**	**MPEG-4 Visual**	**All**	**Mobile**	**Tablet**	**All**	**All Data**
PLCC	0.8108	0.8208	0.8147	0.7527	0.8027	0.7818	0.8881	0.9484	0.9041	0.8525
RMSE	10.2285	9.1770	9.9050	10.0418	8.6199	9.4539	10.4684	7.0315	9.6720	9.9952
SROCC	0.6339	0.7204	0.7276	0.6488	0.7602	0.7725	0.8623	0.9297	0.8893	0.7559
OR	0.0233	0.0000	0.0111	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0042

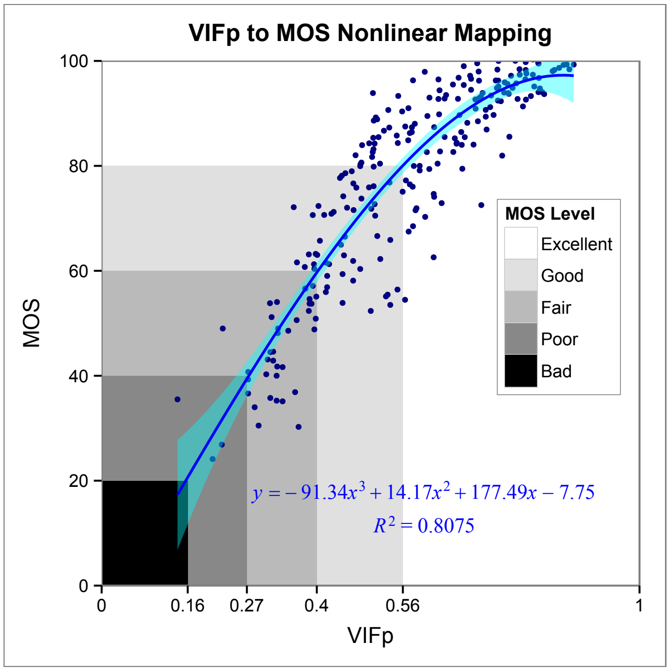

Figure 1. Nonlinear cubic polynomial mapping function between the values of the VIFp metric and the subjective MOS scores for the combined data of the ECVQ, EVVQ and LIVE Mobile databases.

Table VI
MAPPING OF THE VIFP OBJECTIVE VQA METRIC VALUES TO THE SUBJECTIVE MOS SCALE, OBTAINED BASED ON THE COMBINED DATA FROM THE ECVQ, EVVQ AND LIVE MOBILE DATABASES.

MOS	**VIFp Values**
5 (Excellent)	$\geq$0.56
4 (Good)	$\geq$0.40 & $<$0.56
3 (Fair)	$\geq$0.28 & $<$0.40
2 (Poor)	$\geq$0.16 & $<$0.28
1 (Bad)	$<$0.16

Table VII presents the quality estimation performance of the obtained VIFp mapping solution, across the different subsets from the three databases (i.e., H.264 vs. MPEG-4 Visual for the ECVQ and EVVQ databases, Mobile vs. Tablet for the LIVE Mobile database). The quality estimation performance is analysed in terms of accuracy (using PLCC and RMSE), monotonicity (using SROCC) and consistency (using OR). The PLCC results show that across the different databases, the quality estimation accuracy of the VIFp metric's mapping varies from 0.7818 in case of the EVVQ database to 0.9041 in case of the LIVE Mobile database. Across individual subsets the estimation accuracy is better for the MPEG-4 Visual subset than for the H.264 subset both in case of the ECVQ database (0.8208 vs. 0.8108), and in case of the EVVQ database (0.8027 vs. 0.7527). Moreover the PLCC quality estimation accuracy is better for the tablet subset that the smartphone subset in case of the LIVE Mobile dataset (0.9484 vs. 0.8881). Similar conclusions can also be drawn in terms of the RMSE quality estimation accuracy results, SROCC quality estimation monotonicity results, and the OR quality estimation consistency results.

V. CONCLUSIONS AND FUTURE WORK

Automatic multimedia quality assessment methods with clear interpretations are increasingly needed, as multimedia services are growing at a fast pace and users are becoming more quality-aware. This paper has proposed a novel methodology for mapping the values of full-reference objective VQA metrics to the subjective MOS scale, based on subjective data from public VQA databases.

The performance of six different full-reference objective VQA metrics was compared using data from three recent databases, that provide subjective ratings for multiple compressed test sequences with different content characteristics and resolutions. A mapping solution using the proposed methodology was demonstrated and evaluated for the Visual Information Fidelity pixel domain version (VIFp) full-reference objective VQA metric, which was shown to present the best performance among the compared metrics.

Future work will aim to improve the proposed methodology by considering other mapping functions. The evaluation of the mapping methodology will also be extended for additional objective VQA metrics, and subjective data for test sequences with different resolution, framerate and bitrate values.

REFERENCES

[1] Cisco, "Cisco Visual Networking Index: Global Mobile Data Traffic Forecast Update, 2013–2018," Tech. Rep., Feb. 2014. [Online]. Available: http://www.cisco.com/c/en/us/solutions/service-provider/visual-networking-index-vni/index.html

[2] A.-N. Moldovan, S. Weibelzahl, and C. H. Muntean, “Energy-Aware Mobile Learning: Opportunities and Challenges,” *IEEE Communications Surveys & Tutorials*, vol. 16, no. 1, pp. 234–265, 2014.

[3] A.-N. Moldovan, I. Ghergulescu, S. Weibelzahl, and C. H. Muntean, “User-centered EEG-based Multimedia Quality Assessment,” in *8th IEEE International Symposium on Broadband Multimedia Systems and Broadcasting (BMSB 2013)*. London, UK: IEEE, 2013.

[4] S. Chikkerur, V. Sundaram, M. Reisslein, and L. J. Karam, “Objective Video Quality Assessment Methods: A Classification, Review, and Performance Comparison,” *IEEE Transactions on Broadcasting*, vol. 57, no. 2, pp. 165–182, 2011.

[5] A.-N. Moldovan and C. H. Muntean, “Subjective Assessment of BitDetect - A Mechanism for Energy-Aware Multimedia Content Adaptation,” *IEEE Transactions on Broadcasting*, vol. 58, no. 3, pp. 480–492, 2012.

[6] S. Winkler, “Analysis of Public Image and Video Databases for Quality Assessment,” *IEEE Journal of Selected Topics in Signal Processing*, vol. 6, no. 6, pp. 616 –625, 2012.

[7] G.-M. Muntean, P. Perry, and L. Murphy, “Objective and Subjective Evaluation of QOAS Video Streaming over Broadband Networks,” *IEEE Transactions on Network and Service Management*, vol. 2, no. 1, pp. 19–28, 2005.

[8] ITU-R, “BT.500-12: Methodology for the subjective assessment of the quality of television pictures,” International Telecommunication Union, Geneva, Switzerland, Tech. Rep., Sep. 2009. [Online]. Available: http://www.itu.int/rec/R-REC-BT.500-12-200909-I/en

[9] ITU-T, “P.910: Subjective video quality assessment methods for multimedia applications,” International Telecommunication Union, Geneva, Switzerland, Tech. Rep., Apr. 2008. [Online]. Available: http://www.itu.int/rec/T-REC-P.910-200804-I/en

[10] Q. Huynh-Thu, M. N. Garcia, F. Speranza, P. Corriveau, and A. Raake, “Study of Rating Scales for Subjective Quality Assessment of High-Definition Video,” *IEEE Transactions on Broadcasting*, vol. 57, no. 1, pp. 1–14, 2011.

[11] M. H. Pinson, N. Staelens, and A. Webster, “The history of video quality model validation,” in *2013 IEEE 15th International Workshop on Multimedia Signal Processing (MMSP)*, 2013, pp. 458–463.

[12] G.-M. Muntean, G. Ghinea, and T. N. Sheehan, “Region of Interest-Based Adaptive Multimedia Streaming Scheme,” *IEEE Transactions on Broadcasting*, vol. 54, no. 2, pp. 296–303, 2008.

[13] M. Kennedy, H. Venkataraman, and G.-M. Muntean, “Battery and Stream-Aware Adaptive Multimedia Delivery for Wireless Devices,” in *2010 IEEE 35th Conference on Local Computer Networks (LCN), IEEE International Workshop on Performance and Management of Wireless and Mobile Networks (P2MNET)*, Denver, Colorado, USA, 2010, pp. 843–846.

[14] S. S. Hemami and A. R. Reibman, “No-reference image and video quality estimation: Applications and human-motivated design,” *Signal Processing: Image Communication*, vol. 25, no. 7, pp. 469–481, 2010.

[15] I. Gunawan and M. Ghanbari, “Efficient Reduced-Reference Video Quality Meter,” *IEEE Transactions on Broadcasting*, vol. 54, no. 3, pp. 669–679, 2008.

[16] Z. Wang and A. C. Bovik, “Mean squared error: Love it or leave it? A new look at Signal Fidelity Measures,” *IEEE Signal Processing Magazine*, vol. 26, no. 1, pp. 98–117, 2009.

[17] Z. Wang, L. Lu, and A. C. Bovik, “Video Quality Assessment Based on Structural Distortion Measurement,” *Signal Processing: Image Communication, Special Issue on “Objective Video Quality Metrics”*, vol. 19, no. 2, p. 121–132, 2004.

[18] R. Trestian, A.-N. Moldovan, C. H. Muntean, O. Ormond, and G.-M. Muntean, “Quality Utility modelling for multimedia applications for Android Mobile devices,” in *7th IEEE International Symposium on Broadband Multimedia Systems and Broadcasting (BMSB 2012)*. Seoul, Korea: IEEE, 2012.

[19] C.-H. Ke, C.-K. Shieh, W.-S. Hwang, and A. Ziviani, “An Evaluation Framework for More Realistic Simulations of MPEG Video Transmission,” *Journal of Information Science and Engineering*, vol. 24, no. 2, pp. 425–440, 2008.

[20] T. Zinner, O. Abboud, O. Hohlfeld, T. Hossfeld, and P. Tran-Gia, “Towards QoE Management for Scalable Video Streaming,” in *Proceedings of the 21th ITC Specialist Seminar on Multimedia Applications - Traffic, Performance and QoE*, Miyazaki, Japan, 2010.

[21] VQEG, “Final Report on the Validation of Video Quality Models for High Definition Video Content,” Video Quality Experts Group, Tech. Rep., Jun. 2010. [Online]. Available: http://www.its.bldrdoc.gov/vqeg/projects/hdtv/hdtv.aspx

[22] U. Engelke, M. Kusuma, H.-J. Zepernick, and M. Caldera, “Reduced-reference metric design for objective perceptual quality assessment in wireless imaging,” *Signal Processing: Image Communication*, vol. 24, no. 7, pp. 525–547, 2009.

[23] M. Vranješ, S. Rimac-Drlje, and K. Grgić, “Review of objective video quality metrics and performance comparison using different databases,” *Signal Processing: Image Communication*, vol. 28, no. 1, pp. 1–19, 2013.

[24] A. K. Moorthy, L. K. Choi, A. C. Bovik, and G. de Veciana, “Video Quality Assessment on Mobile Devices: Subjective, Behavioral and Objective Studies,” *IEEE Journal of Selected Topics in Signal Processing*, vol. 6, no. 6, pp. 652 –671, 2012.

[25] P. Hanhart, “VQMT: Video Quality Measurement Tool | Multimedia Signal Processing Group (MMSPG),” Mar. 2013. [Online]. Available: http://mmspg.epfl.ch/vqmt

[26] Z. Wang, E. P. Simoncelli, and A. C. Bovik, “Multiscale Structural Similarity for Image Quality Assessment,” in *Conference Record of the Thirty-Seventh Asilomar Conference on Signals, Systems and Computers, 2004*, vol. 2, 2003, pp. 1398–1402.

[27] H. Sheikh and A. Bovik, “Image Information and Visual Quality,” *IEEE Transactions on Image Processing*, vol. 15, no. 2, pp. 430–444, 2006.

[28] K. Egiazarian, J. Astola, N. Ponomarenko, V. Lukin, F. Battisti, and M. Carli, “A New Full-reference Quality Metrics Based on HVS,” in *Proceedings of the Second International Workshop on Video Processing and Quality Metrics*, vol. 4, Scottsdale, Arizona, U.S.A., 2006. [Online]. Available: http://enpub.fulton.asu.edu/resp/vpqm/vpqm2006/papers06/270.pdf

[29] N. Ponomarenko, F. Silvestri, K. Egiazarian, M. Carli, J. Astola, and V. Lukin, “On between-coefficient contrast masking of DCT basis functions,” in *Proceedings of the Third International Workshop on Video Processing and Quality Metrics*, vol. 4, 2007. [Online]. Available: http://ponomarenko.info/vpqm07_p.pdf