Perceptual Video Quality Estimation By Regression With Myopic Experts

Vasileios Giotsas*, Nikos Deligiannis† ‡, Pam Fisher* and Yiannis Andreopoulos‡
*BAFTA Research
British Academy of Film and Television Arts (BAFTA), London, UK
Email: v.giotsas@bafta.org, pamf@bafta.org
†Department of Electronic and Informatics (ETRO)
Vrije Universiteit Brussel (VUB), Brussels, Belgium
Email: ndeligia@etro.vub.ac.be
‡Department of Electronic and Electrical Engineering
University College London (UCL), London, UK
Email: i.andreopoulos@ucl.ac.uk

***Abstract*—Objective video quality metrics can be viewed as "myopic" expert systems that focus on particular aspects of visual information in video, such as image edges or motion parameters. We conjecture that the combination of many such high-level metrics leads to statistically-significant improvement in the prediction of reference-based perceptual video quality in comparison to each individual metric. To examine this hypothesis in a systematic and rigorous manner, we use: *(i)* the LIVE and the EPFL/PoliMi databases that provide the difference mean opinion scores (DMOS) for several video sequences under encoding and packet-loss errors; *(ii)* ten well-known metrics that range from mean-squared error based criteria to sophisticated visual quality estimators; *(iii)* five variants of regression-based supervised learning. For 400 experimental trials with random (non-overlapping) estimation and prediction subsets taken from both databases, we show that the best of our regression methods: *(i)* leads to *statistically-significant* improvement against the best individual metrics for DMOS prediction for more than 97% of the experimental trials; *(ii)* is *statistically-equivalent* to the performance of humans rating the video quality for 36.75% of the experiments with the EPFL/PoliMi database. On the contrary, no single metric achieves such statistical equivalence to human raters in any of the experimental trials.**

***Index Terms*—perceptual video quality, objective metrics, supervised learning**

I. Introduction

Perceptual video quality is pertinent to all video storage and distribution systems that perform lossy operations to a reference (original) video, such as encoding, transcoding and lossy video streaming over IP or wireless networks. The perceptual (a.k.a. subjective) quality of a video sequence is quantified via controlled tests with human raters. Specifically, mean and variance-based normalization of the "difference scores" (i.e., the score given by each rater to the reference—or undistorted—video minus the score given to the distorted version) [1] is performed. The normalized difference scores are then scaled to the range $[0, 100]$ and, after outlier rejection, averaging over all human subjects that rated the particular video is performed to create its *difference mean opinion score* (DMOS) [1]. The standard deviation of the normalized–and–scaled difference scores for each video is also kept to indicate the divergence of opinions of human raters for the particular video content.

This work was funded in part by Innovate UK, project REVQUAL – Resolving Visual Quality for Media (101855).

A. Visual Quality Metrics

Several visual quality metrics have been developed to estimate perceptual video quality. Their performance is quantified based on their statistical correlation to the DMOS of each video within test video databases [1]. These automated metrics go beyond the well-known peak signal-to-noise ratio (PSNR) or structural similarity index metric (SSIM) [2]. Some of the most successful ones are summarized below:

- Multiscale-SSIM (MS-SSIM), an extension of the SSIM paradigm for still images [3], that has been shown to outperform the SSIM index and many other still image quality assessment algorithms [4]. Similar to PSNR and SSIM, the MS-SSIM index is extended to video by applying it frame-by-frame on the luminance component of each video frame and the overall MS-SSIM index for the video is computed as the average of the frame level quality scores [1].
- Visual Information Fidelity (VIF) [5], an image information measure that quantifies the information that is present in the reference image and how much of this reference information can be extracted from the distorted image.
- P-HVS (PSNR - Human Visual System) [6] and P-HVSM [7], which are two weighted versions of PSNR that take into account contrast sensitivity in the pixel [6] and discrete cosine transform domain [7], respectively.
- MOtion-based Video Integrity Evaluation (MOVIE) index in its temporal, spatial and aggregate forms, a.k.a. T-MOVIE, S-MOVIE and MOVIE [8], that perform an optical flow estimation and a Gabor spatial decomposition in order to extract temporal and spatial quality indexes against a reference video.

- Video Quality Model (VQM) [9], a video quality assessment algorithm adopted by ANSI and ITU-T as a standard metric for visual quality assessment; VQM performs spatio-temporal calibration in the input video and then extracts perception-based features (based on spatio-temporal activity detection in short video segments) and computes and combines together video quality parameters to produce a single metric for visual quality.

Previous work [1], [10], [11], [12] focuses on comparisons of such metrics on publicly-available databases of original and distorted video content, i.e., the LIVE [1] and the EPFL-PoliMi [11]. These two databases contain a mixture of four different distortion types: MPEG-2 compression, H.264 compression, simulated transmission of H.264 compressed bitstreams through error-prone IP networks and through error-prone wireless networks. Therefore, they are now becoming the *de-facto* standard for perceptual video quality assessment as they circumvent certain issues with Video Quality Experts Group (VQEG) studies, namely their use of outdated or interlaced content, their poor perceptual separation of videos and the fact that the videos were not made publicly available [12], [1].

B. Related Work and Paper Contribution

The use of *low-level* feature vectors (e.g., color, 2D cepstrum, weighted pixel differencing, spatial decomposition coefficients, etc.) and machine learning for perceptual quality estimation of still images is established by recent studies [13], [14], [15], [16], [17], [18]. In addition, recent work examined the adaptive fusion of multiple objective quality metrics according to distortion types [19]. The efficacy of objective metrics in predicting the visual quality degradation incurred by certain distortion types has also been studied by Reibman, Barkowsky, Vu *et al* [20], [21], [22], [23], as well as by the ITU-T VQEG/JEG committees and others [24], [25]. These important results provide insight on the influence of certain types of degradations in the visual quality of images and video and may help in devising new objective metrics for visual quality assessment. However, in many scenarios encountered in entertainment and creative industries, it is not common to have reliable *a-priori* knowledge about the degradation imposed on video during the production and distribution toolchains. Therefore, there is significant interest in methods that devise advanced visual quality estimation approaches without such knowledge.

In this paper, rather than manually examining the difference in performance between the various quality metrics for different degradation types and test conditions, we conjecture for the first time that *high-level* perceptual video quality metrics can be viewed jointly as a group of "myopic" experts, each focusing on particular aspects of video distortion perceived by human viewers. Therefore, we hypothesize that their combination will lead to improved prediction of experimentally-derived DMOS values of video content. To this end, we opt for the use of *regression-based* rather than *classification-based* supervised learning methods, since the DMOS values span most of the $[0, 100]$ interval. Importantly, due to the large number of repetitions of each test using non-overlapping training and testing subsets selected at random (Monte-Carlo robustness), the conclusions of our study are not strongly dependent on the exact nature and size of the utilized video databases, or on the training and testing subsets themselves. In specific, the results show that our approach allows for *statistically-significant improvement* against the best of the individual metrics and that all metrics contribute to this goal, albeit not equally. Importantly, up to 36.75% of our experimental DMOS-prediction trials with the EPFL/PoliMi database were found to be *statistically equivalent* to the performance of human video-quality raters (a.k.a. the "optimal" prediction model). This is a significant result given that no individual metrics can achieve such statistical equivalence in any test, even when their values are *fitted to the entire set* of DMOS values via logistic scaling [1], [26].

II. Regression With Myopic Experts

In order to estimate the regression parameters, we separate each video database into two equal-size, non-overlapping, subsets: the *estimation* and *prediction* subsets, with $1 \leq j_{\mathrm{e}} \leq J_{\mathrm{e}}$ and $1 \leq j_{\mathrm{p}} \leq J_{\mathrm{p}}$ the indices within each subset and $J_{\mathrm{e}}+J_{\mathrm{p}} = J_{\mathrm{total}}$ the total number of test videos in each database. By randomly shuffling the video indexing, we can generate T_{trial} *experimental trials* with non-overlapping, estimation and prediction subsets. This removes the bias introduced from the usage of a specific estimation and prediction subset and allows us to draw conclusions on the efficacy of our approach independent of the particular video content used for training and testing.

We denote by $m_{\mathrm{e},i,j_{\mathrm{e}}}$ (resp. $m_{\mathrm{p},i,j_{\mathrm{p}}}$) the ith visual metric value for the j_{e}th (resp. j_{p}th) video, with the metric numbering, $1 \leq i \leq 10$, following the order they were mentioned in the previous section (i.e., order of appearance of underlined names of metrics) and $1 \leq j_{\mathrm{e}} \leq J_{\mathrm{e}}$ (resp. $1 \leq j_{\mathrm{p}} \leq J_{\mathrm{p}}$) the index of each video in the estimation (resp. prediction) subset of each database. The ensemble of metrics for the j_{e}th (resp. j_{p}th) video comprises the 10×1 vector $\mathbf{m}_{\mathrm{e},j_{\mathrm{e}}}$ (resp. $\mathbf{m}_{\mathrm{p},j_{\mathrm{p}}}$). The DMOS value and standard deviation of the normalized-and-scaled difference scores for the j_{e}th (resp. j_{p}th) video are denoted by $\mathrm{d}_{\mathrm{e},j_{\mathrm{e}}}$ and $\mathrm{s}_{\mathrm{e},j_{\mathrm{e}}}$ (resp. $\mathrm{d}_{\mathrm{p},j_{\mathrm{p}}}$ and $\mathrm{s}_{\mathrm{p},j_{\mathrm{p}}}$), and are taken from the database results.

For the tth trial, $1 \leq t \leq T_{\mathrm{trial}}$, each approach starts from a random parameter-estimation subset of DMOS and metrics values:

$$\mathbf{d}_{\mathrm{e}}(t) = \left[\mathrm{d}_{\mathrm{e},1}(t) \quad \cdots \quad \mathrm{d}_{\mathrm{e},J_{\mathrm{e}}}(t)\right], \tag{1}$$

and

$$\mathbf{M}_{\mathrm{e}}(t) = \left[\mathbf{m}_{\mathrm{e},1}(t) \quad \cdots \quad \mathbf{m}_{\mathrm{e},J_{\mathrm{e}}}(t)\right]. \tag{2}$$

First, as carried out in previous studies [1], the four-parameter logistic scaling function (recommended by VQEG [27], [1]) is used for each individual metric, with non-linear fitting carried out using the estimation DMOS and metrics' values ($\mathbf{d}_{\mathrm{e}}(t)$ and $\mathbf{M}_{\mathrm{e}}(t)$) and the `nlinfit` function of Matlab. The parameters of the logistic function are kept for each trial t and are used to logistically scale the corresponding metrics of the prediction subset. We then estimate the 1×11

regression vector, $\mathbf{c}_{\text{method}}(t)$, with each of the proposed regression methods, in order to approximate the DMOS values of the estimation subset via

$$\widehat{\mathbf{d}}_{\text{e}}(t) = \left[\widehat{\text{d}}_{\text{e},1}(t) \quad \cdots \quad \widehat{\text{d}}_{\text{e},J_{\text{e}}}(t)\right] = \mathbf{c}_{\text{method}}(t) \begin{bmatrix} \mathbf{M}_{\text{e}}(t) \\ \mathbf{1} \end{bmatrix}, \tag{3}$$

with $\mathbf{1} = [1 \quad \cdots \quad 1]$ the $1 \times J_{\text{e}}$ vector of ones. For each trial t, the aim of each regression method is to minimize the L_z norm error $\left\|\mathbf{d}_{\text{e}}(t) - \widehat{\mathbf{d}}_{\text{e}}(t)\right\|_z$, $z \in \{1, 2\}$, in the *estimation* subset with the expectation that this will also minimize the error between the predicted DMOS

$$\widehat{\mathbf{d}}_{\text{p}}(t) = \left[\widehat{\text{d}}_{\text{p},1}(t) \quad \cdots \quad \widehat{\text{d}}_{\text{p},J_{\text{p}}}(t)\right] \tag{4}$$

and the ground-truth DMOS

$$\mathbf{d}_{\text{p}}(t) = \left[\text{d}_{\text{p},1}(t) \quad \cdots \quad \text{d}_{\text{p},J_{\text{p}}}(t)\right] \tag{5}$$

in the *prediction* subset.

A. Ordinary Least Squares Regression

Starting with ordinary least squares (OLS) regression that minimizes the L_2 norm of the DMOS prediction error, we estimate $\mathbf{c}_{\text{OLS}}(t)$ for each trial t via the estimation subset:

$$\mathbf{c}_{\text{OLS}}(t) = \left[\left(\mathbf{M}_{\text{e}}(t)\left[\mathbf{M}_{\text{e}}(t)\right]^{\text{T}}\right)^{-1}\mathbf{M}_{\text{e}}(t)\left[\mathbf{d}_{\text{e}}(t)\right]^{\text{T}}\right]^{\text{T}}, \tag{6}$$

with superscript T denoting matrix or vector transposition. Once calculated by (6), $\mathbf{c}_{\text{OLS}}(t)$ can be used in conjunction with the metrics for the prediction subset, $\mathbf{M}_{\text{p}}(t) = \left[\mathbf{m}_{\text{p},1}(t) \quad \cdots \quad \mathbf{m}_{\text{p},J_{\text{p}}}(t)\right]$, for the prediction of $\mathbf{d}_{\text{p}}(t)$.

B. L_1 Regression

Instead of minimizing the L_2 norm of the DMOS prediction error, we can instead minimize the L_1 norm via L_1 regression [28]. This can be done via the following iterative process:

1) The initial regression coefficients, $\mathbf{c}_{\text{L1}}^{(0)}(t)$, are calculated via (6) and we set $i = 1$.
2) Compute the $1 \times J_{\text{e}}$ vector

$$\mathbf{w}^{(i)} = \left|\mathbf{d}_{\text{e}}(t) - \mathbf{c}_{\text{L1}}^{(0)}(t)\begin{bmatrix}\mathbf{M}_{\text{e}}(t) \\ \mathbf{1}\end{bmatrix}\right|^{-1}. \tag{7}$$

3) Calculate the updated regression coefficients by ($\text{diag}(\mathbf{w})$ is the diagonal matrix containing weights $\mathbf{w}$):

$$\mathbf{c}_{\text{L1}}^{(i)}(t) = \left[\left(\mathbf{M}_{\text{e}}(t)\,\text{diag}\left(\mathbf{w}^{(i)}\right)\left[\mathbf{M}_{\text{e}}(t)\right]^{\text{T}}\right)^{-1} \times \mathbf{M}_{\text{e}}(t)\,\text{diag}\left(\mathbf{w}^{(i)}\right)\left[\mathbf{d}_{\text{e}}(t)\right]^{\text{T}}\right]^{\text{T}}. \tag{8}$$

4) If

$$\left\|\mathbf{c}_{\text{L1}}^{(i)}(t) - \mathbf{c}_{\text{L1}}^{(i-1)}(t)\right\|_2 \leq e_{\text{thres}}, \tag{9}$$

with e_{thres} a predetermined threshold, then stop; else, set $i \leftarrow i + 1$ and go to Step 2.

This process is guaranteed to converge after a finite number of steps [28] and the final coefficients, $\mathbf{c}_{\text{L1}}(t)$, are used in conjunction with $\mathbf{M}_{\text{p}}(t)$ to predict the DMOS values of the prediction subset, $\mathbf{d}_{\text{p}}(t)$.

C. Variational Bayesian Linear Regression

Alternative approaches to classical multiple linear regression models of (6) and (8) can be constructed based on a Bayesian framework. Unless based on an overly simplistic parametrization, however, exact inference in Bayesian regression models is analytically intractable. This problem can be overcome using methods for approximate inference [29], [30] to construct a framework for variational Bayesian linear (VBL) regression [31], [32], [30].

We consider OLS regression with a shrinkage prior on the regression coefficients [29]. For each trial t, $1 \leq t \leq T_{\text{trial}}$, we wish to infer on the coefficients $\mathbf{c}_{\text{VBL}}(t)$ their precision $\alpha(t)$ and the noise precision $\lambda(t)$. Since there is no analytic expression for the posterior probability density function (PDF) $p(\mathbf{c}_{\text{VBL}}(t), \alpha(t), \lambda(t) | \mathbf{d}_{\text{e}}(t))$, we seek a variational approximation [29], [32] of this posterior PDF starting with the product of the three marginal PDFs of $\mathbf{c}_{\text{VBL}}(t)$, $\alpha(t)$ and $\lambda(t)$ and monitoring the approximation of the lower bound of $\log p(\mathbf{c}_{\text{VBL}}(t), \alpha(t), \lambda(t) | \mathbf{d}_{\text{e}}(t))$ via an iterative process [32], [31]. Pseudocode for VBL regression is given in Algorithm 1 of Ting *et al.* [31]. For our experiments, the VBL regression was realized via the TAPAS library [30].

D. Regression Trees

Another alternative approach we study for the DMOS prediction is regression trees (RT) [28]. Starting with the parameter estimation subset, we examine all possible binary splits on a set of pre-established predictors. A split is selected according to the mean squared error (MSE) of the prediction. If the split leads to a child node having too few observations (less than a predetermined constraint), a split that minimizes the MSE subject to the constraint on observations is selected. This process is repeated recursively for the two child nodes. For each node, the splitting is terminated when: *(i)* the observed response in this node drops below the MSE for the observed response in the entire data multiplied by a predetermined tolerance level on the quadratic error per node, or *(ii)* there are fewer observations in this node than a predetermined value. This method was implemented using the `RegressionTree` class of Matlab.

E. Partial Least Squares Regression

The final approach we study is used when one needs to find predictors for DMOS values that are the most relevant to the observed data. In particular, for the tth trial, $1 \leq t \leq T_{\text{trial}}$, partial least squares regression searches for a set of components (called latent vectors) that performs a simultaneous decomposition of $\mathbf{d}_{\text{e}}(t)$ and $\mathbf{M}_{\text{e}}(t)$ with the constraint that these components explain as much as possible of the covariance between $\mathbf{d}_{\text{e}}(t)$ and $\mathbf{M}_{\text{e}}(t)$ [28]. It is then followed by a classical regression step where the decomposition of $\mathbf{M}_{\text{e}}(t)$ is used to predict $\mathbf{d}_{\text{e}}(t)$. The derived regression

coefficients are then used in the prediction phase of each trial t to predict $\mathbf{d}_{\mathrm{p}}(t)$ from $\mathbf{M}_{\mathrm{p}}(t)$. PLS was implemented using the `plsregress` function of Matlab. We opted to retain six PLS components, which forms a good compromise between dimensionality reduction and predictive power of the regression.

III. Experiments

In our experiments, we use the $J_{\mathrm{e}} = J_{\mathrm{p}} = \frac{J_{\mathrm{total}}}{2}$ video sequences for estimation and prediction ($J_{\mathrm{total}} = 150$ and $J_{\mathrm{total}} = 144$ for the LIVE and the EPFL/PoliMi databases, respectively) and perform $T_{\mathrm{trial}} = 400$ independent trials. For presentation consistency, the EPFL/PoliMi database data were scaled to the $[0, 100]$ range employed by the LIVE database [1]. Moreover, the standard deviation values of the EPFL/PoliMi database were derived from the reported 95% confidence intervals. We measure the efficiency of each approach via: *(i)* the mean absolute error of the DMOS prediction

$$M_{\mathrm{method}} = \frac{1}{T_{\mathrm{trial}} \times J_{\mathrm{p}}} \sum_{t=1}^{T_{\mathrm{trial}}} \left\| \widehat{\mathbf{d}}_{\mathrm{p}}(t) - \mathbf{d}_{\mathrm{p}}(t) \right\|_1 ; \qquad (10)$$

(ii) the percentage of times each DMOS prediction, $\forall j_{\mathrm{p}} \in \{1, J_{\mathrm{p}}\} : \widehat{\mathrm{d}}_{j_{\mathrm{p}}}(t)$, falls within

$$\left[\mathrm{d}_{j_{\mathrm{p}}}(t) - \mathrm{s}_{j_{\mathrm{p}}}(t), \mathrm{d}_{j_{\mathrm{p}}}(t) + \mathrm{s}_{j_{\mathrm{p}}}(t) \right],$$

i.e., within one standard deviation from the corresponding experimental measurement; *(iii)* the average adjusted R^2 correlation coefficient [28], which is computed over all T_{trial} tests by

$$R^2_{\mathrm{method}} = 1 - \frac{J_{\mathrm{p}} - 1}{T_{\mathrm{trial}}(J_{\mathrm{p}} - w_{\mathrm{method}} - 1)} \times \sum_{t=1}^{T_{\mathrm{trial}}} \frac{\left\| \widehat{\mathbf{d}}_{\mathrm{p}}(t) - \mathbf{d}_{\mathrm{p}}(t) \right\|_2^2}{\sum_{j_{\mathrm{p}}=1}^{J_{\mathrm{p}}} \left(\mathrm{d}_{j_{\mathrm{p}}}(t) - \frac{1}{J_{\mathrm{p}}} \sum_{j_{\mathrm{p}}=1}^{J_{\mathrm{p}}} \mathrm{d}_{j_{\mathrm{p}}}(t) \right)^2} \qquad (11)$$

with w_{method} the total number of coefficients (regressors) of each model. Specifically, $w_{\mathrm{method}} = 0$ for each single-metric method and $w_{\mathrm{method}} = 11$ for all regression methods. The adjustment of R^2_{method} according to w_{method} is done to take into account the use of multiple regressors and avoid spuriously increasing of R^2_{method} by overfitting [28].

Table I presents the results for all methods. With the exception of regression trees, the proposed regression methods bring 9% to 34% improvement in the mean adjusted R^2_{method} value in comparison to the best of the individual metrics. By comparing OLS, L_1, VBL and PLS regression to the best individual metrics (i.e., VQM and S-MOVIE), we observe 9% to 19% increase in the percentage of predicted DMOS values that fall within one standard deviation from the experimental DMOS values against the best individual metrics. In addition, the mean absolute error of the DMOS prediction is decreased by 27% to 35%. Importantly, we have confirmed that removing *any* of the utilized metrics from the regression (even some of the worst performing ones), the adjusted R^2_{method} values of all

TABLE I. Mean absolute error, percentage of results within one standard deviation of the experimental DMOS and average adjusted R^2_{method} value, over all T_{trial} trials.

Database	LIVE [1]			EPFL/PoliMi [11]		
Single-metric Method	M_{method}	% in 1 std	R^2_{method}	M_{method}	% in 1 std	R^2_{method}
PSNR	7.94	65.79	0.22	12.92	40.03	0.53
SSIM	8.03	65.82	0.19	15.49	31.81	0.38
MS-SSIM	6.02	78.43	0.48	7.88	59.79	0.83
VIF	7.97	66.80	0.18	14.07	40.01	0.44
P-HVS	7.38	70.21	0.32	10.70	47.37	0.68
P-HVSM	6.95	73.06	0.41	8.56	55.62	0.80
S-MOVIE	6.72	74.98	0.42	**7.39**	**61.25**	**0.85**
T-MOVIE	7.12	70.31	0.37	9.15	48.02	0.79
MOVIE	6.86	72.91	0.41	8.60	54.76	0.80
VQM	**5.82**	**83.92**	**0.56**	8.50	52.92	0.81
Proposed Method	M_{method}	% in 1 std	R^2_{method}	M_{method}	% in 1 std	R^2_{method}
OLS	4.30	93.14	0.77	4.81	**79.84**	0.94
L_1	**4.26**	93.27	0.77	5.05	77.31	**0.96**
VBL	4.41	92.63	0.75	**4.81**	79.49	0.94
RT	5.56	84.19	0.61	7.58	59.15	0.85
PLS	**4.26**	**93.31**	**0.78**	5.04	78.81	0.93

three regression methods decrease between 3% to 35%. This indicates that all metrics are indeed contributing to the final DMOS prediction, albeit not to the same extent.

To examine whether these improvements are statistically significant, we performed F-tests (at 1% false-rejection probability) between all regression methods and the best single-metric methods, i.e., VQM and S-MOVIE. We calculate the related F-statistic for each trial t of each case by

$$F_{\mathrm{method,metric}}(t) = \left(\frac{J_{\mathrm{p}}}{w_{\mathrm{method}}} - 1 \right) \left(\frac{\mathrm{SSR}_{\mathrm{metric}}(t)}{\mathrm{SSR}_{\mathrm{method}}(t)} - 1 \right), \qquad (12)$$

with: $\mathrm{SSR}_{\mathrm{metric}}(t)$ the sum of the squared residual (SSR) error of each single-metric method at the tth experimental trial; $\mathrm{SSR}_{\mathrm{method}}(t)$ the SSR error of each regression-based method at the tth trial; and $w_{\mathrm{method}} = 11$ the degrees of freedom of each regression method. The "null" hypothesis of each F-test is that the DMOS prediction improvement via regression is *not* statistically significant, i.e.,

$$F_{\mathrm{method,metric}}(t) \leq \mathcal{F}^{-1}(0.99, w_{\mathrm{method}}, J_{\mathrm{p}} - w_{\mathrm{method}}), \qquad (13)$$

with $\mathcal{F}^{-1}(1 - a, b, c)$ the value of the inverse $\mathcal{F}$ distribution (*F-threshold*) at false-rejection probability a with (b, c) degrees of freedom [28]. The results are given in Table II. The $F_{\mathrm{method,metric}}(t)$ values of the best regression methods (OLS, VBL and PLS) are higher than the threshold F-ratio for 97% to 100% of experimental trials. Therefore, the null hypothesis is rejected for more than 97% of our experiments, i.e., OLS, VBL and PLS regression lead to *statistically-significant improvement against all single-metric* DMOS prediction methods for the vast majority of experimental trials.

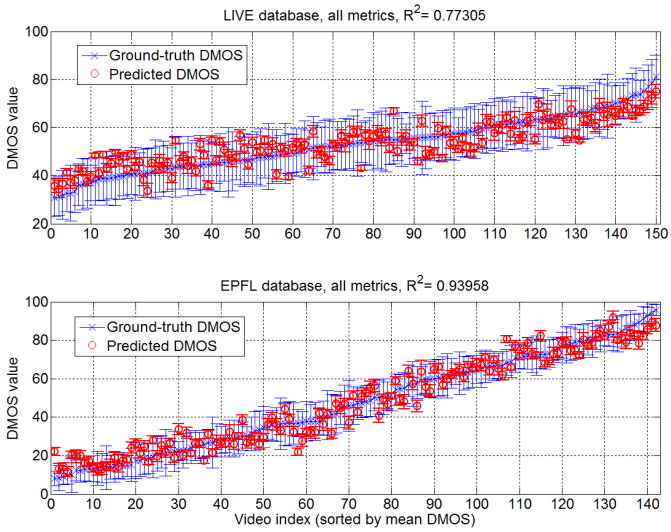

Fig. 1. Ground-truth and predicted DMOS values (with the standard deviation of individual ratings in bars) over all trials for the proposed OLS regression.

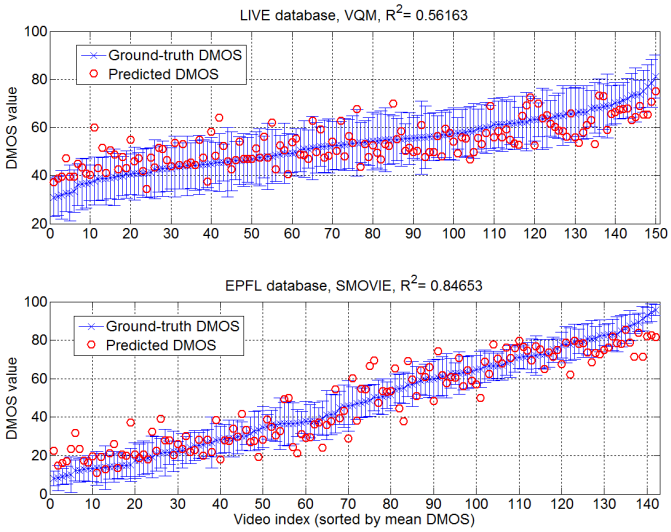

Fig. 2. Ground-truth DMOS values (with the standard deviation of individual ratings in bars) and predicted DMOS by the VQM and S-MOVIE metrics.

TABLE II. AVERAGE $F_{\text{METHOD,METRIC}}(t)$ VALUES (OVER ALL TRIALS t) OF OLS, L_1 AND VBL REGRESSION AGAINST THE VQM AND S-MOVIE METRICS AND, IN BRACKETS, PERCENTAGE OF THE EXPERIMENTAL TRIALS THAT WERE FOUND TO BE ABOVE THE THRESHOLD F-RATIO AT 1% FALSE-REJECTION PROBABILITY.

Database	LIVE [1]		EPFL/PoliMi [11]	
Method/Metric	VQM	S-MOVIE	VQM	S-MOVIE
OLS	8.71 [100%]	13.80 [100%]	15.16 [100%]	10.76 [97%]
L_1	8.44 [99%]	13.43 [99%]	12.93 [100%]	9.14 [90%]
VBL	7.90 [98%]	12.72 [98%]	15.18 [100%]	10.95 [97%]
RT	1.94 [31%]	1.94 [26%]	1.64 [31%]	1.64 [26%]
PLS	7.08 [99%]	7.08 [99%]	11.00 [100%]	11.00 [100%]
F-ratio	2.54		2.56	

To visually illustrate the improvement in the DMOS prediction against the best single metrics, we order all video sequences in terms of their DMOS and present: *(i)* the ground-truth DMOS and standard deviation of difference scores of human raters; *(ii)* the DMOS predicted by the proposed OLS regression; *(iii)* the DMOS predicted by the best single-metric methods. The results are given in Fig. 1 and Fig. 2. It is shown that, while the S-MOVIE and VQM metrics do not predict several of the low and high DMOS values well, the proposed OLS regression provides for significantly more reliable predictions across the entire range of DMOS values.

The standard deviations in Fig. 1 and Fig. 2 illustrate the expected deviations between the experimental DMOS per video and the individual quality ratings given by each human rater to each video. These deviations cannot be predicted by any objective model [1]. Therefore, for each experimental trial t, the optimal model [1], i.e., the ensemble of ground-truth human ratings, has SSR error $\text{SSR}_{\text{optimal}}(t)$, that corresponds to the sum of squared residual error between individual subjective ratings and the video DMOS [1]. Such SSR errors can also be calculated between individual subjective ratings and the best regression-based models (denoted by $\text{SSR}_{\text{model,subj}}(t)$).

Focusing on the EPFL/PoliMi database where the full ensemble of human ratings is publicly available, for each experimental trial t we performed an F-test (at 1% false-rejection probability) to determine whether our regression-based approaches can be deemed to be *statistically equivalent* to the optimal model. That is, we check for how many trials the following holds:

$$\frac{\text{SSR}_{\text{model,subj}}(t)}{\text{SSR}_{\text{optimal}}(t)} \leq \mathcal{F}^{-1}(0.99, J_{\text{p}}, 40 \times J_{\text{p}}), \qquad (14)$$

where 40 corresponds to the number of individual human raters of the database. We found that this occurred in: *(i)* 35% of trials for OLS regression; *(ii)* 28.75% of the trials for L_1 regression; *(iii)* 36.75% of the trials for VBL regression; *(iv)* 1.04% of the trials for RT; *(v)* 32.18% of the trials for PLS regression. On the contrary, and as reported in previous studies [1], [26], we never found this to be the case for any of the trials with any of the individual metrics. To the best of our knowledge, this is the first time a DMOS prediction approach exhibits statistical equivalence to the optimal (a.k.a. ground-truth) model for a substantial percentage of experimental trials.

IV. CONCLUSIONS

By viewing multiple high-level visual quality metrics as myopic experts, we proposed and validated for the first time their combination for the prediction of difference mean opinion scores (DMOS) of video sequences. Five regression-based methods and two publicly-available databases have been used for our experiments. Given the Monte-Carlo based training and testing, we are able to draw rigorous conclusions on the efficacy of our approach that are independent of the particular video content used for training and testing. In particular, it is found that significant improvement in the DMOS prediction accuracy is offered by four out of the five regression methods against the best of the single metrics. In addition, all five methods are shown to outperform individual metrics. For four out of five regression methods, this improvement is found to be statistically significant for more than 97% of the randomized training and testing. This constitutes a solid validation of our hypothesis that combining multiple high-level visual quality

metrics is beneficial for DMOS prediction. Moreover, variational Bayesian regression is found to be statistically equivalent to the performance of human raters of video quality for 36.75% of the experimental trials with the EPFL/PoliMi database. This may form a significant step towards the ultimate goal of creating an expert system for reference-based prediction of perceptual video quality that is indistinguishable from the ensemble of human ratings. Future work can validate our approach in further datasets, as well as against methods that perform selective fusion of metrics according to distortion types, noting however that our approach does not require knowledge or training conditional to distortion types.

REFERENCES

[1] K. Seshadrinathan, R. Soundararajan, A. C. Bovik, and L. K. Cormack, "Study of subjective and objective quality assessment of video," *IEEE Trans. Image Process.*, vol. 19, no. 6, pp. 1427–1441, 2010.

[2] Z. Wang, A. C. Bovik, H. R. Sheikh, and E. P. Simoncelli, "Image quality assessment: From error visibility to structural similarity," *IEEE Trans. Image Process.*, vol. 13, no. 4, pp. 600–612, 2004.

[3] Z. Wang, E. P. Simoncelli, and A. C. Bovik, "Multiscale structural similarity for image quality assessment," in *IEEE 37th Asilomar Conf. Sig., Syst. and Comp., 2003*, vol. 2. IEEE, 2003, pp. 1398–1402.

[4] H. R. Sheikh, M. F. Sabir, and A. C. Bovik, "A statistical evaluation of recent full reference image quality assessment algorithms," *IEEE Trans. Image Process.*, vol. 15, no. 11, pp. 3440–3451, 2006.

[5] H. R. Sheikh and A. C. Bovik, "Image information and visual quality," *IEEE Trans. Image Process.*, vol. 15, no. 2, pp. 430–444, 2006.

[6] K. Egiazarian, J. Astola, N. Ponomarenko, V. Lukin, F. Battisti, and M. Carli, "New full-reference quality metrics based on hvs," in *Proc. 2nd Int. Workshop Video Process. and Qual. Metr.*, vol. 4, 2006.

[7] N. Ponomarenko, F. Silvestri, K. Egiazarian, M. Carli, J. Astola, and V. Lukin, "On between-coefficient contrast masking of dct basis functions," in *Proc. 3rd Int. Workshop Video Process. and Qual. Metr.*, vol. 4, 2007.

[8] K. Seshadrinathan and A. C. Bovik, "Motion tuned spatio-temporal quality assessment of natural videos," *IEEE Trans. Image Process.*, vol. 19, no. 2, pp. 335–350, 2010.

[9] M. H. Pinson and S. Wolf, "A new standardized method for objectively measuring video quality," *IEEE Trans. Broadcast.*, vol. 50, no. 3, pp. 312–322, 2004.

[10] K. Seshadrinathan, R. Soundararajan, A. Bovik, and L. Cormack, "Study of subjective and objective quality assessment of video," *IEEE Trans. Image Process.*, vol. 19, no. 6, pp. 1427–1441, 2010.

[11] F. De Simone, M. Naccari, M. Tagliasacchi, F. Dufaux, S. Tubaro, and T. Ebrahimi, "Subjective assessment of h. 264/avc video sequences transmitted over a noisy channel," in *IEEE Int. Workshop Qual. of Multim. Exper., QoMEx 2009*. IEEE, 2009, pp. 204–209.

[12] A. C. Bovik *et al.*, "Live image quality assessment database," *website: http://live.ece.utexas.edu/research/quality*, 2003–2014.

[13] M. Narwaria and W. Lin, "SVD-based quality metric for image and video using machine learning," *IEEE Trans. Syst., Man, and Cybern., Part B*, vol. 42, no. 2, pp. 347–364, 2012.

[14] C. Charrier, O. Lézoray, and G. Lebrun, "Machine learning to design full-reference image quality assessment algorithm," *J. Sig. Process.: Image Comm.*, vol. 27, no. 3, pp. 209–219, 2012.

[15] P. Peng and Z.-N. Li, "A mixture of experts approach to multi-strategy image quality assessment," in *Proc. Image Anal. and Rec., ICIAR 2012*. Springer, 2012, pp. 123–130.

[16] P. Gastaldo and J. A. Redi, "Machine learning solutions for objective visual quality assessment," in *Proc. 6th Int. Workshop on Video Process. and Qual. Metrics for Consumer Electr. (VPQM-12)*, 2012.

[17] M. Narwaria, W. Lin, and A. E. Cetin, "Scalable image quality assessment with 2d mel-cepstrum and machine learning approach," *J. Patt. Rec.*, vol. 45, no. 1, pp. 299–313, 2012.

[18] P. Peng and Z. Li, "Image quality assessment based on distortion-aware decision fusion," in *Proc. Intel. Sci. and Intel. Data Eng., IScIDE 2011*. Springer, 2012, pp. 644–651.

[19] A. Barri, A. Dooms, B. Jansen, and P. Schelkens, "A locally adaptive system for the fusion of objective quality measures," *IEEE Trans. on Image Processing*, vol. 23, no. 6, pp. 2446–2458, June 2014.

[20] 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, pp. 013 016–013 016, 2014.

[21] A. R. Reibman, "A strategy to jointly test image quality estimators subjectively," in *Proc. 19th IEEE Int. Conf. on Image Process. (ICIP)*. IEEE, 2012, pp. 1501–1504.

[22] F. M. Ciaramello and A. R. Reibman, "Systematic stress testing of image quality estimators," in *Proc. 18th IEEE Int. Conf. on Image Process. (ICIP)*. IEEE, 2011, pp. 3101–3104.

[23] M. Barkowsky, E. Masala, G. Van Wallendael, K. BRUNNSTRÖM, N. Staelens, and P. Le Callet, "Objective video quality assessmentâtowards large scale video database enhanced model development," *IEICE Trans. on Comm.*, vol. 98, no. 1, pp. 2–11, 2015.

[24] I. Tutorial, "Objective perceptual assessment of video quality: full reference television," *ITU-T Telecommunication Standardization Bureau*, 2004.

[25] S. Chikkerur, V. Sundaram, M. Reisslein, and L. J. Karam, "Objective video quality assessment methods: A classification, review, and performance comparison," *IEEE Trans. on Broadcasting*, vol. 57, no. 2, pp. 165–182, 2011.

[26] K. Brunnstrom, D. Hands, F. Speranza, and A. Webster, "VQEG validation and ITU standardization of objective perceptual video quality metrics [Standards in a Nutshell]," *IEEE Sig. Process. Mag.*, vol. 26, no. 3, pp. 96–101, 2009.

[27] R. Streijl, S. Winkler, and D. Hands, "Perceptual quality measurement: Towards a more efficient process for validating objective models [standards in a nutshell]," *IEEE Signal Process. Mag.*, vol. 27, no. 4, pp. 136–140, July 2010.

[28] D. Birkes and Y. Dodge, *Alternative methods of regression*. John Wiley & Sons, 2011, vol. 190.

[29] C. M. Bishop and N. M. Nasrabadi, *Pattern recognition and machine learning*. springer New York, 2006, vol. 1.

[30] C. Mathys, J. Daunizeau, K. J. Friston, and K. E. Stephan, "A bayesian foundation for individual learning under uncertainty," *Frontiers in human neuroscience*, vol. 5, 2011.

[31] J.-A. Ting, A. D'Souza, K. Yamamoto, T. Yoshioka, D. Hoffman, S. Kakei, L. Sergio, J. Kalaska, M. Kawato, P. Strick *et al.*, "Variational bayesian least squares: an application to brain–machine interface data," *Neural Networks*, vol. 21, no. 8, pp. 1112–1131, 2008.

[32] A. Armagan, "Variational bridge regression," in *Proc. Int. Conf. on Artificial Intell. and Stat.*, 2009, pp. 17–24.