IEICE TRANS. INF. & SYST., VOL.E96–D, NO.3 MARCH 2013

PAPER

A Reduced-Reference Video Quality Assessment Method Based on the Activity-Difference of DCT Coefficients

Wyllian B. da SILVA[†a)], ***Student Member***, **Keiko V. O. FONSECA**[†b)], ***Member***, ***and*** **Alexandre de A. P. POHL**[†c)], ***Nonmember***

SUMMARY A simple and efficient reduced-reference video quality assessment method based on the activity-difference of DCT coefficients is proposed. The method provides better accuracy, monotonicity, and consistent predictions than the PSNR full-reference metric and comparable results with the full-reference SSIM. It also shows an improved performance to a similar VQ technique based on the calculation of the pixel luminance differences performed in the spatial-domain.

key words: *DCT, reduced-reference metric, video quality assessment*

1. Introduction

In recent years there has been a growing development of objective video quality assessment methods for multimedia processing, such as acquisition, compression, transmission, restoration, storage, segmentation, and presentation [1]–[3]. Especially, in digital systems quality assessment of streaming services and digital TV broadcasting these issues are currently discussed in the literature [4]–[6].

The evaluation of video impairments can be performed using objective methods, which can be classified as signal fidelity measures and Perceptual Visual Quality Metrics (PVQMs) [2], [3]. The conventional objective methods, such as the Mean Absolute Error (MAE), the Mean Square Error (MSE), the Signal-to-Noise Ratio (SNR) and the Peak Signal-to-Noise Ratio (PSNR) are understood as signal fidelity measures [7]. On the other hand, PVQMs can be orderd into two categories [8]: the signal-driven approach and the vision-based modeling. The first involves the evaluation of luminance/color distortion, statistical features, structural similarity and visual artifacts, such as blockiness and blurring. The second category includes psychophysical properties and physiological knowledge, such as the Contrast Sensitivity Function (CSF), luminance adaptation, temporal, spatial and color decomposition, and several masking effects [2]. PVQMs can still be classified as double-ended and single-ended metrics. The single-ended metric, also known as No-Reference (NR), needs only the processed video [2], [7], [9], [10]. The double-ended metric can be of two kinds: Reduced-Reference (RR) and Full-Reference (FR). FR metrics require full information of the video source on the measurement site. On the other hand, the RR metrics require access only to portions of the source information (video reference), forming a set of RR features that can facilitate the evaluation process. While FR metrics require the presence of the video source and NR metrics are still immature, RR techniques can provide operators with a functional tool for evaluation of the video quality delivered to customers.

Previous development of RR methods applied to image and video can be found in the literature. For instance, Tao *et al.* [11] reported a method based on the contourlet domain that decomposes images and then extracts features to mimic the multichannel structure of the Human Visual System (HVS). Their method also incorporates the CSF, which is applied to weigh coefficients and the Weber's law of Just Noticeable Difference (JND) to produce a noticeable variation in sensory experience. Ma *et al.* in [12] developed a technique based on the statistical modeling of the Discrete Cosine Transform (DCT) coefficient distributions, exploiting the identical nature of the distributions between adjacent subbands and the coefficients into a three-level tree using the Generalized Gaussian Density (GGD) function. Concerning the evaluation of video, Gunawan and Ghanbari [13], [14] described a method where a discriminative analysis of the harmonic strength computed from the edge-detected frame is employed to create harmonics of gain and loss information. Hewage and Martini [15] proposed a metric for 3D video (encoded in H.264/AVC format), which uses PSNR and is based on the edge detection. Their method uses the depth map information difference between the sender and receiver side obtained with the Sobel filtering. Finally, Yamada *et al.* [16] reported a simple RR metric based on PSNR, named as Video Quality (VQ), which consists in calculating the absolute difference between the luminance and the mean luminance values in a 16×16 pixel block.

In this work we propose a Reduced-Reference Video Quality Assessment (RRVQA) method that can be applied both in digital TV systems and in video streaming services, where the video quality prediction is calculated based on the frequency domain activity-difference between the reference and the received video. The idea behind the improved technique is to explore the variation of coefficients between the sender and receiver side, which occur due to errors or data distortions in the high frequency components. The method has the advantage of easy of implementation, because it involves the video post-processing through the DCT trans-

Manuscript received May 28, 2012.
Manuscript revised October 31, 2012.

[†]The authors are with the Graduate Program on Electrical Engineering and Applied Computer Science – CPGEI, Federal University of Technology – Paraná, UTFPR, Curitiba, Paraná, Brazil.
a) E-mail: wyllianbs@gmail.com
b) E-mail: keiko@utfpr.edu.br
c) E-mail: pohl@utfpr.edu.br
DOI: 10.1587/transinf.E96.D.708

Copyright © 2013 The Institute of Electronics, Information and Communication Engineers

form, which can eventually be incorporated in the decoding step. Besides, it requires a reduced number of bits, which is an important feature of RR techniques. For instance, in [17], [18] 8 bits per pixel are required and in [19] and [20] 5 bits and 1 bit per pixel, respectively, are needed. On the other hand, our approach requires about 19 bits per macroblock to represent the information from the video source.

We demonstrate that the method enables a more accurate prediction of the video quality when compared with results reported in [16], whose approach is based on the spatial domain analysis and where the adjacent 8×8 pixel blocks present a high degree of correlation. By using the DCT transform, it uncorrelates the information in the 8×8 blocks and its neighbors. For testing the method, videos from the LIVE database [21] are used, which include videos distorted by MPEG-2 and H.264 compression, error-prone wireless networks, and IP networks.

The paper is organized as follows. Section 2 describes the reduced-reference quality assessment method based on the frequency domain. In Sect. 3 details of the experiments, the used database and the quality calibration are presented. Section 4 shows the performance comparison between the proposed RRVQA algorithm and PSNR, Structural SIMilarity index (SSIM) and the VQ [16], followed by the discussion of the results and the conclusion in Sect. 5.

2. The Reduced-Reference Assessment Method for Video Quality

It's well known that high frequency components are responsible for details in a frame, for which the HVS is not sensitive [22]. Lossy compression processes (e.g., MPEG-2 or H.264) suppress these components and produce artifacts, such as blurring and blocking. Figures 1 (a) and 1 (b) show, respectively, the original and degraded frame of the video named Tractor, where degradation was inflicted by the loss of frames in simulated IP transmission. The zoom box in Fig. 1 (b) shows, as an example, the generated artifact due to packet loss. Given that the high frequency components are more susceptible to changes, the proposed method explores the variation of the DCT coefficients (particularly, the AC coefficients) between the original and distorted frames at the sender and receiver sides.

This is most evident when one observes the behavior of the 256 first DCT coefficient values for the original and degraded frames of different videos in Fig. 2. The higher peaks correspond to the DC coefficients and have a higher energy (low frequency). On the other hand, the much smaller peaks correspond to the AC coefficients. Figure 2 (a) shows the behavior of DC and AC coefficient values for the original Tractor video and Fig. 2 (b) to 2 (e) the corresponding values for degraded videos due to error-prone wireless networks, H.264 compression, simulated transmission of H.264 compressed bitstreams over error-prone IP networks, and MPEG-2 compression, respectively. One sees that the DC and AC values are severely changed when distortions occurs depending on the video degradation. This is particularly the case for videos degraded by the H.264 compression and by the transmission over error-prone Wireless networks, when changes in both DC and AC coefficients are more evident (see Figs. 2 (b) and 2 (c)).

(a)

(b)

Fig. 1 Frame of the Tractor video from the LIVE database. (a) Original frame. (b) Frame obtained by the simulated transmission of H.264 compressed bitstreams through error-prone IP networks, where PSNR = 24.4113 dB, VQ = 37.3565 dB, and RRVQA = 18.0989 dB.

Base on the behavior of DC and AC coefficients the proposed assessment technique employs the two-dimensional DCT transform [23], which uses the DC and AC coefficients of a $\tau \times \tau$ macroblock, formed by multiple 8×8 blocks.

The coefficients in each macroblock are represented by the variable $coef_p$ (coefficients in each $\tau \times \tau$ macroblock, including the $\frac{\tau \times \tau}{64}$ DC coefficients and also the $\tau \times \tau - \frac{\tau \times \tau}{64}$ AC coefficients) while the DC coefficients absolute average (the mean of the DC coefficients) is represented by the $\overline{DC}$ variable. Thus, the absolute activity-difference for the macroblock j, with $\tau \times \tau$ resolution, given by $(Actf_j)$, is calculated as

$$Actf_j = \frac{1}{\tau \times \tau} \sum_{p=1}^{\tau \times \tau} \left| coef_p - \overline{DC} \right|, \tag{1}$$

where $\overline{DC}$ is the average of $\frac{\tau \times \tau}{64}$ DC coefficients, the denominator is equal to 64 due to the DCT transform that operates on 8×8 blocks. The $\overline{DC}$ value is calculated as

$$\overline{DC} = \frac{64}{\tau \times \tau} \sum_{k=1}^{\frac{\tau \times \tau}{64}} |DC_k| . \tag{2}$$

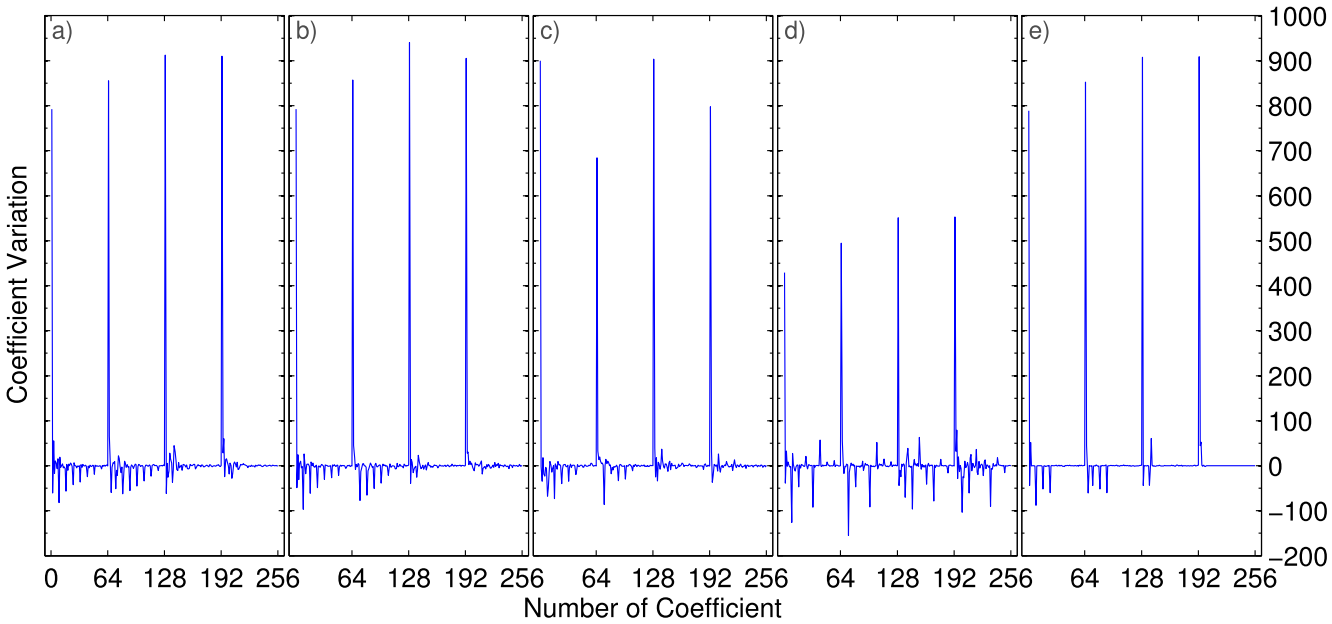

Fig. 2 Signal frequency representation of the first 256 DCT coefficients of the Tractor frame. (a) Original frame in the frequency-domain using the DCT transform. (b) Error-prone wireless networks. (c) H.264 compression. (d) IP loss frame by simulated transmission of H.264 compressed bitstreams through error-prone IP networks in the frequency domain using the DCT transform. (e) MPEG-2 compression.

It is the parameter $Actf_j$ from a frame of the original video that is transmitted by the RR method over the digital TV channel or over the Internet by means of the Transport Stream (TS). Between 8 and 10 bits are required to represent each ($Actf_j$) coefficient (or 1 feature per $\tau \times \tau$ coefficients in a macroblock), while 11-bit floating point (i.e., 8 bits for mantissa and 3 bits for exponent) are required to represent the maximum value of $\overline{DC}$.

The number of bits required to represent each parameter can be further reduced if lossless compression methods are used. At the receiver side the equivalent parameter of the degraded frame is calculated also using Eq. (1). This way, Eq. (3) computes the square of the difference between the activity-difference frequency ($ActfS_j$) on the sender side and that on the receiver side ($Actf R_j$). This difference is named the Square Error (SEf_j) per macroblock j in the frequency-domain, give as

$$SEf_j = \left(ActfS_j - Actf R_j\right)^2. \tag{3}$$

As a matter of comparison, the square errors calculated in the frequency domain (SEf) and those in the space domain (SEs), whose method is described in [16], are given in Fig. 3 for the same frame of Fig. 1 (b). One sees that the SEf values are higher for the frequency domain case, evidencing the higher sensitivity of the AC components to changes. This behavior extends to all frames of the tested videos in this work.

The proposed RRVQA method further requires the computation of the MSE on frequency domain, which represents the average of (SEf_j) for all macroblocks in a frame

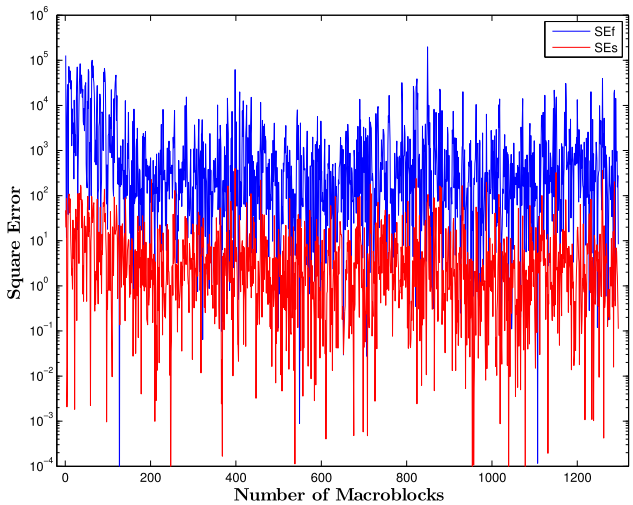

Fig. 3 Behavior of the square error in the frequency-domain (SEf) and the space-domain (SEs) for degradation through simulated transmission over error-prone IP networks.

i, given as

$$MSEf_i = \frac{1}{M}\sum_{j=1}^{M} SEf_j, \tag{4}$$

where M is the number of macroblocks per frame.

Finally, the calculation of the RRVQA index is calculated as

$$RRVQA = \frac{1}{N}\sum_{i=1}^{N} 10 \times log_{10}\frac{[max\,(cs_i, cr_i)]^2}{MSEf_i}, \tag{5}$$

where N is the number of frames in the video; cs_i and cr_i are the DCT coefficients from the sender and receiver side, respectively. Only the $Actf_j$ calculated by (1) and the maximum value of cs_i are transmitted to the receiver side.

3. Description of the Experiments

In order to test the metric performance against results of subjective evaluation (human scores) statistical methods are employed, such as the Spearman Rank Order Correlation Coefficient (SROCC) and the Pearson Linear Correlation Coefficient (PLCC). VQEG recommendations [24]–[29] suggest the discard of references samples during the assessment, as well as the application of mapping and validation experiments. It also recommends the application of Difference Mean Opinion Scores (DMOS) for FR and RR metrics and Mean Opinion Scores (MOS) for NR metric. Thus, in the experiments, only videos from the database without references were employed. The mapping of values obtained with the objective metric to the DMOS scale is performed using the cubic polynomial function, as explained in Sect. 3.2. The resulting prediction characterizes the correlation between two measures, one based on the objective metric (RRVQA) and the other based on a subjective metric (DMOS).

The performance of other metrics in comparison with the proposed RRVQA is also investigated and the corresponding prediction was calculated for the VQ metric reported in [16] and the full reference metrics PSNR and SSIM. The comparison with other metrics requires, however, that the value range of the full-reference PSNR and VQ lies between 0 and 100 dB. This assumption is required because two identical frames give MSE = 0 and PSNR equal infinite. In order to avoid this inconsistency in the calculations, the 100 dB range is defined as an arbitrary upper bound [30], [31]. This maximum value implies that no differences between the reference and processed frames exist.

3.1 Video Database

The LIVE video quality database is used in the experiments. This database includes 150 videos from 10 reference video contents, as shown in Fig. 4. This database includes distorted videos by MPEG-2 and H.264 compression, error-prone wireless networks, and IP networks [21].

The first seven video sequences (from left to right and from top to bottom) have a frame rate of 25 frames per second (fps), while the remaining three (Mobile and Calendar, Park Run, and Shields) have a frame rate of 50 fps. All video files do not contain headers and have planar YUV 4:2:0 formats, whose resolution is 768×432 pixels. The LIVE video database only contains DMOS subjective samples.

3.2 Quality Calibration

The mapping of the objective score scale into the subjective score scale of DMOS can be performed using either a non-linear logistic function [24] or non-linear polynomial functions, according to the Video Quality Experts Group (VQEG) recommendation [26]. This mapping must provide a simple empirical prediction and not cause an overfitting of data points. In this work, the mapping was performed between DMOS and RRVQA (x) using a cubic polynomial function [25]–[28] defined as

Fig. 4 Videos from the LIVE database.

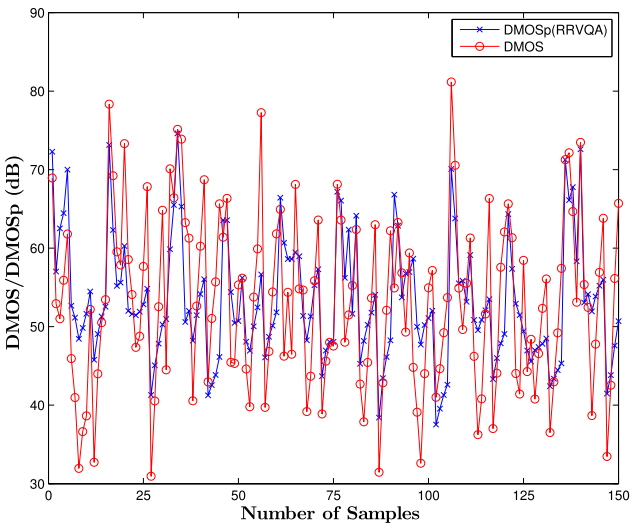

Fig. 5 Scale mapping RRVQA into DMOS with cubic polynomial function for all videos from LIVE database.

$$DMOSp = ax^3 + bx^2 + cx + d. \quad (6)$$

For example, in Fig. 5, the coefficients were found to be $a = 0.0022$; $b = -0.1214$; $c = -0.4285$; and $d = 108.2439$, where mapping was performed using the RRVQA (x) metric.

The mapping showed that the cubic polynomial function is better suited as it does not cause overfitting of data points at the low extreme, as would happen if the monotonic logistic function were employed [32]. This way, PLCC and SROCC were computed after performing the non-linear regression using the cubic polynomial function.

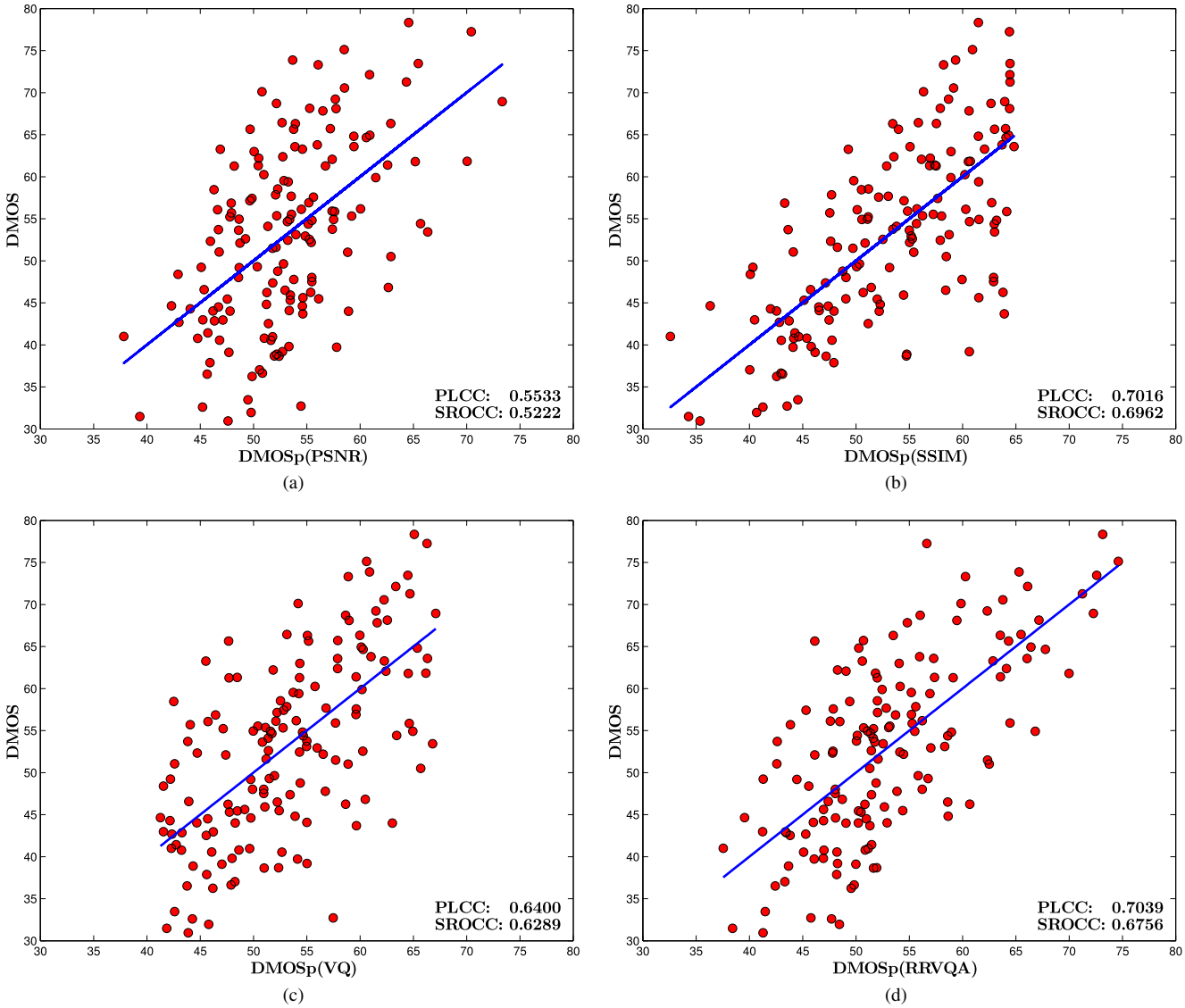

Fig. 6 Correlation between objective metric scale and subjective (DMOS) metric scale with cubic polynomial function for all data of the LIVE video database. (a) PSNR vs. DMOS; (b) SSIM vs. DMOS; (c) VQ [16] vs. DMOS; (d) proposed RRVQA vs. DMOS.

4. Results and Discussion

The experiments were conducted using $\tau = 16$ in Eq. (1) for the proposed method. Figure 6 shows the correlation behavior between the DMOS and PSNR, SSIM, VQ, and RRVQA using the video database from LIVE. The PLCC between DMOS and RRVQA is equal to 0.7039, while that of PSNR, SSIM and VQ are 0.5533, 0.7016 and 0.64, respectively.

The SROCC for RRVQA is 0.6756, while that of PSNR, SSIM, and VQ metrics are 0.5222, 0.6962 and 0.6289, respectively. The visual inspection of Fig. 6 (d), as compared to the others 6 (a), 6 (b) and 6 (c), confirm the numerical results expressed above.

For each video set of the database the PLCC and SROCC coefficients, R-Square, RMSE, 95% confidence limits, statistical significance (ζ), percentage of F-distribution (F_p), outlier ratio (OR), and MAE were computed. Data shown in bold type in Tables 1, 2, 3, 5 and 6 point out to the best score in each content category.

Table 1 compares PSNR and SSIM full-reference metrics with scores obtained by the reduced-reference metrics (VQ [16] and the proposed RRVQA method). Table 2 compares the R-Square between RRVQA and PSNR, SSIM, and VQ and Table 3 presents the RMSE and confidence limits with 95% ($C_{95\%}$). Table 2 also shows the N_f parameter, which represents the number of samples employed in the evaluation. SSIM presents RMSE with 95% confidence interval below 7.24 for IP content, while our RRVQA presents RMSE with 95% confidence interval below 7.24 for H.264, IP, and Wireless contents, as recommended in [33]. The comparison using the F-distribution (ζ) and the percentage of F-distribution (F_p) is shown in Table 4. These measures are interesting because quantify the performance between

Table 1 Comparison of accuracy (PLCC) and monotonicity (SROCC) between the RRVQA method and PSNR, SSIM and VQ metrics for samples without reference.

Content	Full-Reference				Reduced-Reference			
	PSNR		SSIM		VQ [16]		proposed RRVQA	
	PLCC	SROCC	PLCC	SROCC	PLCC	SROCC	PLCC	SROCC
All	0.5533	0.5222	0.7016	**0.6962**	0.6400	0.6289	**0.7039**	0.6756
H.264	0.6080	0.5498	0.6889	0.6841	0.7134	0.6861	**0.7862**	**0.7614**
IP	0.6123	0.5754	**0.7051**	**0.6960**	0.6904	0.5813	0.7045	0.6863
MPEG-2	0.5502	0.5069	**0.6893**	**0.6800**	0.5016	0.4763	0.6270	0.6352
Wireless	0.6453	0.6046	0.6844	0.6795	0.7310	0.7462	**0.8082**	**0.7878**

Table 2 Comparison of R-Square between full-reference and reduced-reference metrics for samples without reference.

Content	N_f	PSNR	SSIM	VQ [16]	proposed RRVQA
All	150	0.3062	0.4922	0.4096	**0.4955**
H.264	40	0.3465	0.5343	0.5089	**0.6181**
IP	30	0.2603	**0.5029**	0.4767	0.4963
MPEG-2	40	0.1808	**0.4547**	0.2516	0.3932
Wireless	40	0.4396	0.4440	0.5343	**0.6532**

Table 3 Comparison of RMSE with confidence interval of 95% between full-reference and reduced-reference metrics.

Content	PSNR	SSIM	VQ [16]	proposed RRVQA
All	9.2681±1.0810	7.9285±0.9247	8.5497±0.9972	**7.9031**±0.9218
H.264	9.2499±2.2906	7.8083±1.9337	8.0187±1.9858	**7.0710**±1.7511
IP	8.6472±2.5916	**7.0889**±2.1246	7.2736±2.1799	7.1356±2.1386
MPEG-2	9.0965±2.2527	**7.4217**±1.8379	8.6942±2.1530	7.8290±1.9388
Wireless	8.1425±2.0164	8.1104±2.0085	7.4223±1.8381	**6.4058**±1.5863

Table 4 Comparison of performance between the RRVQA and PSNR, SSIM, and VQ metrics, respectively, with F-distribution (ζ) and percentage of F-distribution F_p for samples without reference.

Content	PSNR vs. proposed RRVQA		SSIM vs. proposed RRVQA		VQ [16] vs. proposed RRVQA	
	ζ	$F_p(\%)$	ζ	$F_p(\%)$	ζ	$F_p(\%)$
All	1.3754	37.53	1.0064	0.64	1.1703	17.03
H.264	1.7113	71.13	1.2194	21.94	1.2860	28.60
IP	1.4686	46.86	0.9870	-1.30	1.0391	3.91
MPEG-2	1.3500	35.00	0.8987	-10.13	1.2333	23.32
Wireless	1.6157	61.57	1.6030	60.30	1.3425	34.25

Table 5 Outlier ratio (OR) and Mean Absolute Error (MAE) between full-reference and reduced-reference metrics for samples without reference.

Content	Full-Reference				Reduced-Reference			
	PSNR		SSIM		VQ [16]		proposed RRVQA	
	OR	MAE	OR	MAE	OR	MAE	OR	MAE
All	0.0200	7.6151	**0.0067**	**6.2109**	0.0200	6.8941	0.0133	6.3838
H.264	0	7.3675	0	**5.4814**	0	6.3790	**0**	5.6254
IP	0.0333	6.5469	0.0333	**5.3318**	0	5.4664	**0**	5.3550
MPEG-2	0	7.2573	0	**6.1510**	0	6.8751	**0**	6.4896
Wireless	0	6.6011	0.0250	6.2113	0	5.7252	**0**	**4.9422**

Table 6 Comparison of performance between DMOS and DMOSp(RRVQA) using different macroblock sizes.

Content	Macroblock 8×8		Macroblock 16×16		Macroblock 32×32	
	PLCC	SROCC	PLCC	SROCC	PLCC	SROCC
All	**0.7191**	**0.6937**	0.7039	0.6756	0.6801	0.6410
H.264	0.7338	0.7229	**0.7862**	**0.7614**	0.7196	0.6617
IP	0.6977	0.6716	**0.7045**	**0.6863**	0.6805	0.6578
MPEG-2	0.6264	0.6131	**0.6270**	**0.6352**	0.5300	0.4896
Wireless	0.7929	0.7674	0.8082	0.7878	**0.8167**	**0.8135**

two metrics. Results show that the performance of RRVQA is superior than VQ and PSNR, with (F_p) values varying from 17% to 37% (for all video contents). When a particular video content is taken into account, one sees that (F_p) values vary from 35% to 71% in favor of the proposed RRVQA metric in comparison with PSNR.

The comparison between the VQ metric and RRVQA shows, however, mixed performance. For instance, when using the video content the RRVQA metric performs better than VQ regardless of the artifact type affecting the video. Only for IP content in the video database, the RRVQA performance is equivalent to VQ. In addition, the proposed method presents better performance for H.264 and wireless contents when compared to SSIM. However, concerning all video contents RRVQA shows equivalent performance to SSIM.

Table 5 shows the data for the outlier ratio and the MAE concerning the four techniques. RRVQA presents performance equivalent to SSIM for OR and MAE measures. Again, the results of Table 5 make it clear that the RRVQA is suitable for video quality monitoring and assessment.

The influence of the macroblock size on the metric results was also investigated. Table 6 shows the comparison of performance assessed by our method using different macroblock sizes and subjective scores (DMOS). Macroblock sizes of 8×8, 16×16 and 32×32 were used in (1). The data in Table 6 shows that PLCC values obtained with are higher in the analysis of H.264, IP and MPEG-2 contents (0.7862, 0.7045 and 0.6270, respectively). However, higher PLCC scores are obtained using the macroblock size 8×8 for the case of "All" (0.7191) and the macroblock size 32×32 for the case of Wireless contents (0.8167). SROCC scores follow the same trend. PLCC scores can reach a difference of up to 7% as observed from the values in Table 6 depending on the employed size. The advantage of using a higher macroblock size lies on the fact that the method requires less bits per frame, i. e., it reduces drastically the number of frequency-domain activity-difference obtained through Eq. (1) for each macroblock (for instance, only one Actf value per 1024 coefficients is required in the case of 32×32. Instead, four times more Actf values are needed to represent the 16×16 case). However, due to the better performance of the metric when applied to H.264, IP and MPEG-2 contents, the macroblock size 16×16 has been chosen for the calculations through out this work.

The comparison between results of this and other works in the literature [11]–[16] is difficult due to the use of different databases and mapping functions. We use the cubic mapping function, but most works used the logistic mapping function [24], which is a VQEG recommendation that has been overtaken. Therefore, although wishful, it is not possible to compare our results and the ones found elsewhere.

Results obtained with RRVQA can be justified due to the activity-difference of DCT coefficients between sender side and receiver side, as expressed in Eqs. (1) through (5). Figure 7 shows the difference of performance between MSE

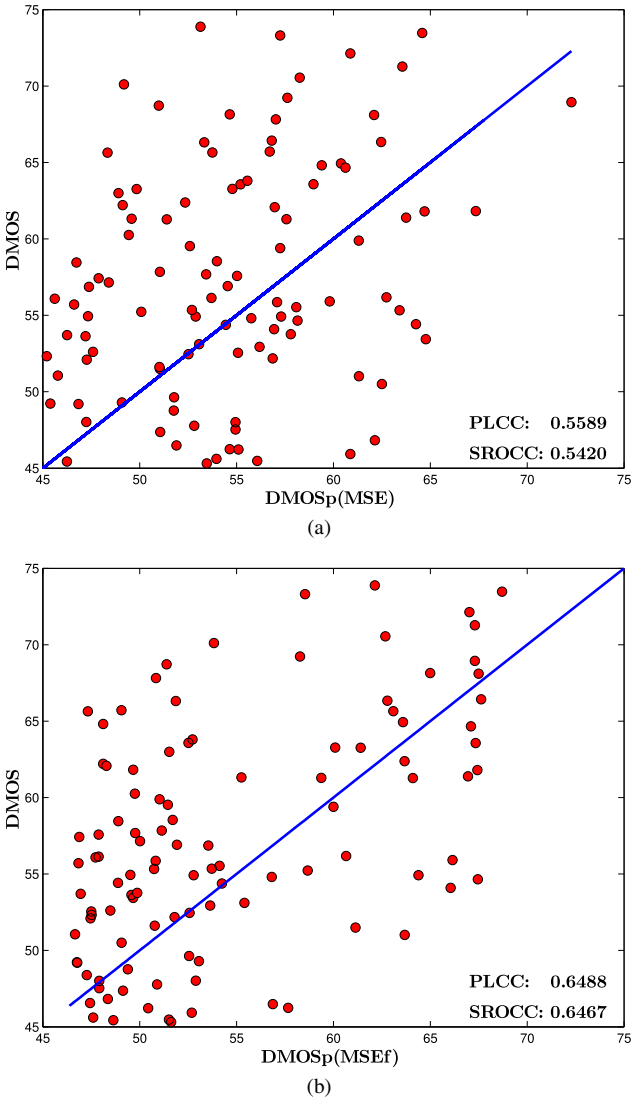

Fig. 7 Comparison between MSE and MSEf on space and frequency domain, respectively. (a) MSE in the spatial domain with PLCC = 0.5589 and SROCC = 0.5420. (b) Proposed MSEf in the frequency domain with PLCC = 06488 and SROCC = 6467.

calculated in the spatial domain and MSEf calculated in the frequency domain, both employed in the denominator of Eq. (5). The sole visual inspection of Figs. 7 (a) and 7 (b) indicates that the approach based on the frequency domain produces better results. PLCC and SROCC for MSE is 0.5589 and 0.5420, respectively, while that for MSEf is 0.6488 and 0.6467.

Our method, although simple, is efficient because it requires low-complexity implementation and it can be incorporated into encoding and decoding processes. In this paper no compression is applied to the $Actf_j$ data that contains the features required by the algorithm. The features are extracted on the receiver side, where the Transport Stream (TS) is demultiplexed and decoded, with 1 feature per 256 DCT coefficients of each 16×16 macroblock. The decoded video may include errors caused by distortions from the channel, attenuation, and compression process or packet

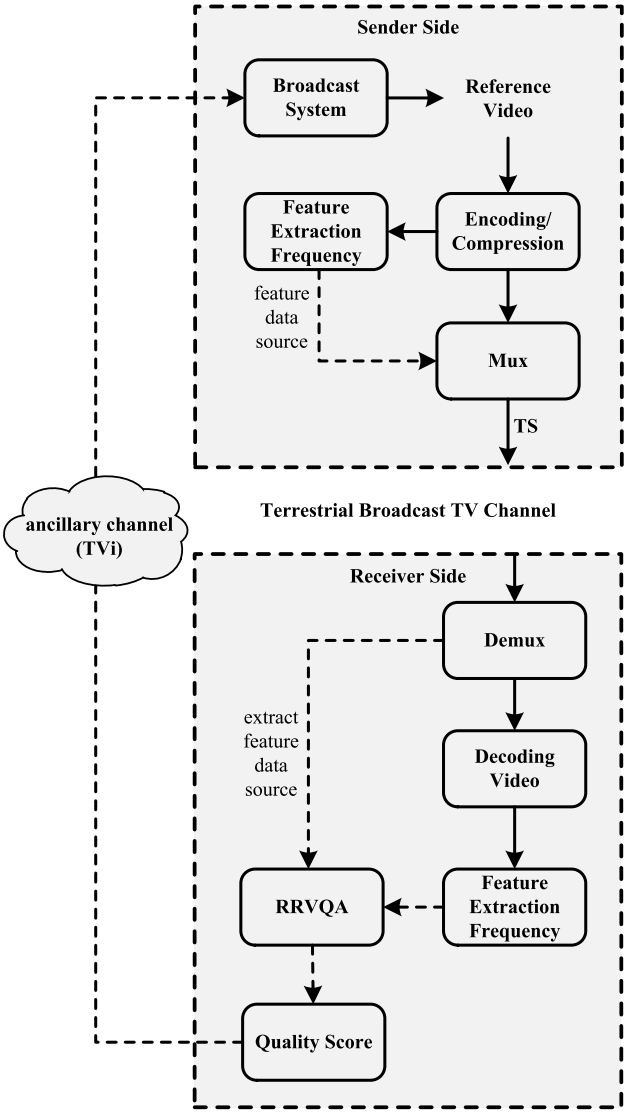

Fig. 8 Framework for video quality-estimation model for digital TV systems based on RRVQA.

loss. Degradations such as blocking and blurring effects may appear causing an unpleasant visual perception to viewers.

Figure 8 shows a framework for video quality assessment in a digital TV system in which the proposed RRVQA algorithm is inserted. It can be applied on the decoded video, creating a quality score that can be sent back to the broadcast system via ancillary channel with TV interactivity (TVi) for further analysis and eventual corrections.

The sender side performs the video compression and encoding process, for instance, using the MPEG-4 Part 10 (H.264) standard. After this step, the feature extraction is multiplexed into the MPEG-2 Transport Stream, modulated and broadcasted over air. The data stream containing the video extracted features data can still be reduced by lossless compression process or other technique to decrease the data size that is transmitted over the broadcast TV channel. This way, the technique can be used in monitoring the video transmission over erasure channels, such as IP networks, or video quality in a digital TV broadcast system.

5. Conclusion

This paper proposes a simple and efficient reduced-reference method for the video quality estimation based on the activity-difference of DCT coefficients. In the spatial domain each set of adjacent 8×8 pixel blocks is correlated, whereas by using the DCT transform it uncorrelates the information between 8×8 blocks and its neighbors. Although the HVS is not sensitive to high frequency, these components are susceptible to errors in a noisy channel and can be detected by the quality assessment method. Therefore, by using the activity-difference of DC and AC coefficients the RRVQA delivers a more accurate prediction of the video quality when compared with the PSNR and VQ techniques, as the numerical results of Tables 1 to 6 demonstrate. From the data it is observed that the proposed method presents smaller predictions errors and higher accuracy, monotonicity and consistency for either the analysis of particular degradations in video sequences (such as artifacts generated by H.264 compression or originated in wireless networks) or when the analysis of all video degradations is considered. Therefore, the proposed method is suitable for the quality monitoring in video transmission over wireless or IP networks and in digital TV systems.

Acknowledgment

Wyllian Bezerra da Silva thanks CAPES for a PhD scholarship. This work has been supported by the project “Formação de Pessoal Qualificado em Sistemas de Transmissão de TV Digital no Paraná – Processo 23038.23556/2008-16 AUX-PE-RH-TVD 249/2008” supported by CAPES. Authors also thank the Laboratory for Image and Video Engineering from the University of Texas at Austin for using the LIVE database.

References

[1] Z. Wang and A.C. Bovik, Modern image quality assessment, Morgan & Claypool, San Rafael, CA, 2006.
[2] W. Lin and C.C.J. Kuo, “Perceptual visual quality metrics: A survey,” J. Visual Communication and Image Representation, vol.22, no.4, pp.297–312, 2011.
[3] Stefan and Winkler, “Issues in vision modeling for perceptual video quality assessment,” Signal Process., vol.78, no.2, pp.231–252, 1999.
[4] S. Gauss, T. Müller, T. Roll, J. Wünschmann, and A. Rothermel, “Objective video quality assessment of mobile television receivers,” Consumer Electronics (ISCE), 2010 IEEE 14th International Symposium on, pp.1–6, june 2010.
[5] D.Z. Rodriguez and G. Bressan, “Video quality assessments on digital TV and video streaming services using objective metrics,” IEEE (Revista IEEE America Latina) Latin America Transactions, vol.10, no.1, pp.1184–1189, 2012.
[6] V.I. Ponomaryov, T. Herfet, V.V. Lukin, B. Smolka, and V. Zlokolica, “Image and video quality improvement techniques for emerging applications,” EURASIP J. Adv. Sig. Proc., vol.2012,

p.33, 2012.
[7] A. Eskicioglu and P. Fisher, "Image quality measures and their performance," IEEE Trans. Commun., vol.43, no.12, pp.2959–2965, Dec. 1995.
[8] S. Winkler and P. Mohandas, "The evolution of video quality measurement: from PSNR to hybrid metrics," IEEE Trans. Broadcast., vol.54, no.3, pp.660–668, 2008.
[9] R.V. Babu, S. Suresh, and A. Perkis, "No-reference JPEG-image quality assessment using GAP-RBF," Signal Process., vol.87, no.6, pp.1493–1503, 2007.
[10] J. Zhang, T.M. Le, S. Ong, and T.Q. Nguyen, "No-reference image quality assessment using structural activity," Signal Process., vol.91, no.11, pp.2575–2588, 2011.
[11] D. Tao, X. Li, W. Lu, and X. Gao, "Reduced-reference IQA in contourlet domain," IEEE Trans. Syst. Man Cybern. B, Cybern., vol.39, no.6, pp.1623–1627, 2009.
[12] L. Ma, S. Li, F. Zhang, and K.N. Ngan, "Reduced-reference image quality assessment using reorganized DCT-based image representation," IEEE Trans. Multimed., vol.13, no.4, pp.824–829, 2011.
[13] I. Gunawan and M. Ghanbari, "Reduced-reference video quality assessment using discriminative local harmonic strength with motion consideration," IEEE Trans. Circuits Syst. Video Technol., vol.18, no.1, pp.71–83, Jan. 2008.
[14] I. Gunawan and M. Ghanbari, "Efficient reduced-reference video quality meter," IEEE Trans. Broadcast., vol.54, no.3, pp.669–679, Sept. 2008.
[15] C. Hewage and M. Martini, "Reduced-reference quality evaluation for compressed depth maps associated with colour plus depth 3D video," 17th IEEE International Conference on Image Processing (ICIP), pp.4017–4020, Sept. 2010.
[16] T. Yamada, Y. Miyamoto, Y. Senda, and M. Serizawa, "Video-quality estimation based on reduced-reference model employing activity-difference," IEICE Trans. Fundamentals, vol.E92-A, no.12, pp.3284–3290, Dec. 2009.
[17] M. Sendashonga and F. Labeau, "Low complexity image quality assessment using frequency domain transforms," IEEE International Conference on Image Processing, pp.385–388, 2006.
[18] Z. Wang, G. Wu, H.R. Sheikh, E.P. Simoncelli, E.H. Yang, and A.C. Bovik, "Quality-aware images," IEEE Trans. Image Process., vol.15, no.6, pp.1680–1689, 2006.
[19] M.D. Gaubatz, R.A. Ulichney, and D.M. Rouse, "A low-complexity reduced-reference print identification algorithm," 16th IEEE International Conference on Image Processing (ICIP), pp.1289–1292, 2009.
[20] C. Hewage and M.G. Martini, "Reduced-reference quality evaluation for compressed depth maps associated with colour plus depth 3D video," 17th IEEE International Conference on Image Processing (ICIP), pp.4017–4020, 2010.
[21] 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.
[22] J. Li, J. Koivusaari, J. Takala, M. Gabbouj, and H. Chen, "Human visual system based adaptive inter quantization," Multimedia on Mobile Devices 2008, vol.6821, 682106, 2008.
[23] K.R. Rao and P. Yip, Discrete cosine transform: algorithms, advantages, applications, Academic Press Professional, Inc., San Diego, CA, USA, 1990.
[24] Video Quality Experts Group (VQEG), "Final report from the video quality experts group on the validation of objective models of video quality assessment," tech. rep., VQEG, 2000.
[25] Video Quality Experts Group (VQEG), "Final report from the video quality experts group on the validation of objective models video quality assessment, Phase II," tech. rep., VQEG, Aug. 2003.
[26] Video Quality Experts Group (VQEG), "Final report from the video quality experts group on the validation of objective models of multimedia quality assessment, Phase I," tech. rep., VQEG, 2008.
[27] Video Quality Experts Group (VQEG), "Final report from the video quality experts group on the validation of reduced-reference and no-reference objective models for standard definition television, Phase I," tech. rep., VQEG, 2009.
[28] Video Quality Experts Group (VQEG), "Report on the validation of video quality models for high definition video content, Version 2.0," tech. rep., VQEG, 2010.
[29] 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 Signal Processing Magazine, vol.26, no.3, pp.96–101, 2009.
[30] G. Gonçalves, R. Dantas, A. Palhares, J. Kelner, J. Fidalgo, D. Sadok, H. Almeida, M. Berg, and D. Cederholm, "Estimating video quality over ADSL2+ under impulsive line disturbance," in AccessNets, ed. C. Wang, O. Akan, P. Bellavista, J. Cao, F. Dressler, D. Ferrari, M. Gerla, H. Kobayashi, S. Palazzo, S. Sahni, X.S. Shen, M. Stan, J. Xiaohua, A. Zomaya, and G. Coulson, Lecture Notes of the Institute for Computer Sciences, Social Informatics and Telecommunications Engineering, vol.6, pp.381–396, Springer Berlin Heidelberg, 2009.
[31] N. Li, B. Yan, G. Chen, P. Govindaswamy, and J. Wang, "Design and implementation of a sensor-based wireless camera system for continuous monitoring in assistive environments," Personal Ubiquitous Comput., vol.14, pp.499–510, Sept. 2010.
[32] U. Engelke, M. Kusuma, H.J. Zepernick, and M. Caldera, "Reduced-reference metric design for objective perceptual quality assessment in wireless imaging," Image Commun., vol.24, pp.525–547, Aug. 2009.
[33] J. Okamoto, T. Hayashi, A. Takahashi, and T. Kurita, "Proposal for an objective video quality assessment method that takes temporal and spatial information into consideration," Electronics and Communications in Japan (Part I: Communications), vol.89, no.12, pp.97–108, 2006.
[34] M.R. Spiegel and L.J. Stephens, Theory and problems of statistics, 3rd ed., Schaum's Outline Series, McGraw-Hill, New York, 1998.

Appendix: Statistical Evaluation Tools

This section describes the statistical tools employed to assess the metric performance and the statistical significance of results when comparison is made with other methods.

The perceptual significance of a metric is determined by the PLCC index (prediction accuracy). If the correlation coefficient approaches 1, the relationship between the scores of the objective metric and the perceptual quality perceived by the HVS is strongly developed. PLCC is calculated using a set of K data pairs (μ_k, ν_k) that can be quantified as [24]–[28], [34]:

$$PLCC = \frac{\sum_{k=1}^{K} (\mu_k - \overline{\mu})(\nu_k - \overline{\nu})}{\sqrt{\sum_{k=1}^{K} (\mu_k - \overline{\mu})^2} \sqrt{\sum_{k=1}^{K} (\nu_k - \overline{\nu})^2}}, \qquad (A \cdot 1)$$

where μ_k and ν_k are the feature difference and the subjective rating related to the k^{th} frame, respectively; $\overline{\mu}$ and $\overline{\nu}$ are the means of the respective data sets.

Monotonicity is used to quantify if changes in one measure (increase or decrease) is followed by magnitude changes (increase or decrease) with respect to another measure. Monotonicity is quantified by the SROCC and is described as [24]–[28], [34].

$$SROCC = \frac{\sum_{k=1}^{K}(\chi_k - \overline{\chi})(\gamma_k - \overline{\gamma})}{\sqrt{\sum_{k=1}^{K}(\chi_k - \overline{\chi})^2}\sqrt{\sum_{k=1}^{K}(\gamma_k - \overline{\gamma})^2}}, \quad (A\cdot 2)$$

where χ_k and γ_k are the ranks of the predicted and the subjective scores, respectively. In addition, $\overline{\chi}$ and $\overline{\gamma}$ are the midranks of the respective data sets.

The R-Square (R^2) [34] represents the degree of variations in the subjective metric values (DMOS) and is described by the fit technique [32].

$$R^2 = 1 - \frac{SSE}{SST}, \quad (A\cdot 3)$$

where SSE (Sum of Square Errors) express the sum of the squared prediction errors between DMOS and DMOSp. SST (Total Sum of Squares) represents the sum of squared deviations of DMOS. R-Square assumes values in the interval [0, 1], where a good fit approaches 1.

The Root Mean Squared Error (RMSE) of the absolute prediction error $P_{error}(i)$ when applied to FR and RR metrics expresses the standard error of the fitting between DMOS and DMOSp [24]–[28].

$$RMSE = \sqrt{\left(\frac{1}{N_f - d}\right)\sum_{i=1}^{N_f} P_{error}(i)^2}, \quad (A\cdot 4)$$

where N_f is total number of videos in the analysis (discarding the video references) and d is the number of degrees of freedom of the non-linear mapping function [34]. In this paper, as recommended by VQEG [25]–[28], we adopt $d = 4$ for the 3^{rd}-order cubic polynomial mapping function and $P_{error}(i)$ is calculated with the formula:

$$P_{error}(i) = DMOS(i) - DMOSp(i). \quad (A\cdot 5)$$

The RMSE indicates a better fit for values closer to 0. However, it is recommended that RMSE with a 95% confidence interval be less than 7.24, as experimental studies point out in [33]. Using the $\chi^2(N_f - d)$ distribution, the limits of the 95% confidence interval for RMSE are defined as [24]–[28]

$$\frac{RMSE\sqrt{N_f - d}}{\sqrt{\chi^2_{0.025}(N_f - d)}} < RMSE < \frac{RMSE\sqrt{N_f - d}}{\sqrt{\chi^2_{0.975}(N_f - d)}}. \quad (A\cdot 6)$$

And the half width of this confidence range is given as

$$C_{95\%} = \pm\frac{1}{2}\left|\frac{RMSE\sqrt{N_f - d}}{\sqrt{\chi^2_{0.025}(N_f - d)}} - \frac{RMSE\sqrt{N_f - d}}{\sqrt{\chi^2_{0.975}(N_f - d)}}\right|. \quad (A\cdot 7)$$

In addition, another figure of merit within the scope of this work is the statistical significance (ζ). It represents the difference between the RMSE of two objective metrics and is calculated as [34]

$$\zeta = \frac{\left(RMSE_{objective\ metric}\right)^2}{\left(RMSE_{RRVQA}\right)^2}, \quad (A\cdot 8)$$

where $RMSE_{objective\ metric}$ assumes the values obtained with PSNR, SSIM, or VQ [16] metric. The statistical significance has an F-distribution with $n1$ and $n2$ degrees of freedom. VQEG reports [24]–[28] establishes the ζ parameter to be evaluated against the tabulated value $F(0.05, n1, n2)$ that ensures 95% of significance. The $n1$ and $n2$ degrees of freedom are given by $Nf1 - d$ and $Nf2 - d$, respectively, with $Nf1$ and $Nf2$ expressing the total number of samples used for calculating RMSE and d being the number of parameters in the non-linear mapping function. If ζ is higher than the tabulated value $F(0.05, n1, n2)$, then there is a significant difference between metrics in terms of RMSE, otherwise both metrics are considered equivalent. Another way of expressing this comparison is through the percentage of F-distribution (F_p), defined as

$$F_p = (\zeta - 1) \times 100. \quad (A\cdot 9)$$

If ζ is higher than 1.05 (or is higher than 5%), then there is a significant difference between the metrics.

The outlier ratio (OR) is employed for measuring the prediction consistency, which is defined as [24]–[28]

$$OR = \frac{\rho}{N_f}, \quad (A\cdot 10)$$

where N_f is the number of samples and ρ is the total of outliers, whose number is determined by considering samples out of the interval calculated as

$$|P_{error}(i)| > 2\sigma(DMOS)_i, \quad i = 1, \ldots, N_f, \quad (A\cdot 11)$$

where $\sigma(DMOS)$ is the standard deviation error of each sample. In addition, the MAE prediction between DMOS and DMOSp, given as

$$MAE = \frac{1}{N_f}\sum_{i=1}^{N_f}|P_{error}(i)|, \quad (A\cdot 12)$$

is also computed. Both OR and MAE indicate a better consistency and error predictions when their values approaches 0 [32].

Wyllian B. da Silva was born in 1978. He received the B.S. and M.S. degrees in physics and electrical engineering from the Federal University of Uberlândia, Brazil, in 2005 and 2008, respectively. He's a Ph.D. student in the Electrical and Computer Engineering Department at the Federal University of Technology – Paraná (UTFPR). He is currently researching video quality assessment, designing algorithms for video quality evaluation systems.

Keiko V. O. Fonseca (M'98) received the degree in electrical engineering from the Federal University, Paraná, Curitiba, Brazil, in 1985, the M.S. degree in electrical engineering from the State University of Campinas, Campinas, São Paulo, Brazil, in 1988, and the Ph.D. degree in electrical engineering from the Federal University of Santa Catarina, Florianópolis-SC, Brazil, in 1997. Since 1988, she has been with the Electrical Engineering Department, Federal University of Technology (UTFPR), Paraná, where she leads a research group on high-speed networks and real-time systems. Presently, she is a board member of the Computer Engineering Undergraduate Program (UTFPR, Curitiba), member of the IEEE Communication Society, the Brazilian Computer Society and the Institute of Electronics, Information, and Communication Engineers (Japan).

Alexandre de A. P. Pohl received the B.S. and M.S. degree in Physics in 1983 and 1987, respectively, from the State University of Campinas (Unicamp), Brazil, and the Ph.D. degree in electrical engineering in 1994 from the Technical University of Braunschweig, Germany. From 1987 to 1989 he was with the laser research division of the Brazilian Airspace Technical Center, São José dos Campos. From 1995 to 2000 he worked at the telecommunications division of Furukawa, Inc in Brazil. Since 2001 he is with the electrical engineering department of the Federal University of Technology – Paraná (UTFPR), Curitiba, Brazil, where he leads a research group working in the area of optical fiber communications and digital TV systems. He is a member of the Optical Society of America (OSA), the Brazilian Telecommunications Society (SBrT) and the Brazilian Microwave and Optoelectronics Society (SBMO).