No-Reference Video Quality Assessment Method based on the Levenberg-Marquardt Minimization

Wyllian B. Silva and Alexandre A. P. Pohl

***Resumo*— O processamento e a transmissão de vídeo digital não se realiza sem a ocorrência de degradações que impactam sua qualidade. Disto decorre a necessidade de ferramentas que possam avaliar a qualidade do vídeo transmitido. Neste trabalho é apresentada uma técnica de avaliação objetiva sem referência baseada em um método analítico, o qual atribui pesos distintos aos descritores espaço-temporais, cujos valores são obtidos pela solução de um problema de mínimos quadrados não-linear com o uso do algoritmo de Levenberg-Marquardt.**

***Palavras-Chave*—Algoritmo Levenberg-Marquardt, métrica no-reference, avaliação objetiva de qualidade de vídeo.**

***Abstract*— The processing and transmission of digital video are impaired by degradations that impact its quality. Evaluation tools are then required to assess the quality of transmitted video. In this work we present a no-reference assessment metric based on an analytical method, which takes into account distinct weights for the spatial-temporal descriptors and whose values are obtained by the solution of a nonlinear least squares problem using the Levenberg-Marquardt algorithm.**

***Index Terms*—Levenberg-Marquardt algorithm, no-reference metric, objective video quality assessment.**

I. INTRODUCTION

THE demand for high quality digital videos has put an enormous pressure upon Internet Service Providers (ISPs) and broadcast operators in recent years, which has led to an increasing offer of broadband access to customers. On offering video services, providers and broadcasters are also required to assess constantly the quality of videos being delivered, once these are affected by processing and transmission degradations over the network. In such an effort, several methods for video quality assessment have been developed, which can be broadly classified into Full-Reference (FR), Reduced-Reference (RR) and No-Reference (NR) metrics. At the same time, the need for local assessment, particularly ones far away from the network stations, has driven research to focus lately on NR methods, once FR and RR metrics require previous information about the video source, which put additional constraints to the evaluation process [1]. Although the NR metric needs only the received or processed video, the development of a technique that presents a high degree of correlation for all types of video with subjective measures is far from satisfactory. There are still not many works on the NR metric for video assessment in the literature and most of them are based on the impact of the perception of the Human Visual System (HVS) over the distortions or artifacts, such as blocking and blurring. For instance, in [2] the spatial distortion of each frame in a video is calculated using the differences between the corresponding regions of two adjacent frames in the video sequence. The predicted distortion is then weighted according to the temporal activities of the frame sequence. In [3] the NR quality score is obtained from the bit stream without the need for the complete video decoding. Three factors are taken into account: picture distortion caused by quantization, quality degradation due to packet loss and error propagation, and temporal effects based on the perception by the HVS. In another work [4], the quality score is obtained from the coding error estimation computed in the transform domain (Discrete Cosine Transform, DCT), whose coefficients are corrupted by the quantization noise, followed by the perceptual weighting of the error. Moreover, the availability of the corrupted bit stream is assumed for the analysis, but in fact NR metrics do not take into account the additional encryption or processing by third-party decoders, where only the decoded pixel value is available. In these cases, [5] describes an alternative method based on the pattern estimation of lost macroblocks, which assumes the knowledge of the pixels only. This information is then used in a NR quality monitoring system that delivers an estimate of the Mean Square Error (MSE) distortion caused by channel errors.

In this work we present an alternative and simple analytical method, which takes into account spatial and temporal descriptors, such as blurring and blocking artifacts (A, B and Z activity measures), the temporal information, the temporal difference and the average and weighted Mean Absolute Difference (MAD) between successive frames. Such descriptors are weighted by values obtained through the solution of the nonlinear least squares method using the Levenberg-Marquardt (LM) algorithm [6], which takes as input the different video sequences from a database. This way, the proposed No-Reference Video Quality Assessment (NRVQA) requires first a training phase from which the

W. B. Silva and A. A. P. Pohl is with the Graduate Program on Electrical Engineering and Applied Computer Science – CPGEI, Federal University of Technology – Paraná, UTFPR, Curitiba, Paraná, Brazil (e-mail: pohl@utfpr.edu.br).

weights of descriptors are obtained. Once the weights are available, they are used in the analytical model to assess the quality of videos.

The paper is divided as follows. Section II describes the proposed no-reference video quality assessment method. Experimental results and their discussion are presented in Section III, followed by the conclusion in Section IV.

II. No-Reference Assessment Method for Video Quality

The method explores features in the spatial-temporal domain and is based on the detection of artifacts and differences between successive frames of a video sequence. The method is devised and optimized to assess the quality of Dirac, H.264, IP and MPEG-2 streams. Fig. 1 shows a diagram of the proposed technique. Initially, during the training phase, the particular video database (Dirac, H.264, IP, and MPEG-2 subset) is selected with its subjective Difference Mean Opinion Scores (DMOS) normalized between 0 and 1. Other inputs are also selected, such as the analytical expression, where the spatial-temporal descriptors are outlined and their corresponding initial weights (initial guess), described as β parameters. The LM algorithm is then applied to solve the nonlinear least squares problem posed by the analytical expression in order to optimize the values attributed to the different β´s. Once this step is completed, the testing phase can be initiated, where the optimized β's are loaded and the corresponding quality score of the distorted video sequence is computed. The proposed method combines the detectors of blocking and blurring artifacts (spatial features) and temporal features, such as the temporal information (TI) and the MAD. The descriptors of blurring and blocking artifacts are represented by the features A, B, and the Z activity measure, which have been described in [7]. We denote the luminance of a frame as $y(f,i,j)$ and the motion difference feature between the luminance pixel values, at the same space location in subsequent frames, as $m(f,i,j)$, with $i \in [1,M]$ and $j \in [1,N]$, where M is the number of rows and N the number of columns in the frame. The temporal information is expressed as

$$TI = \frac{1}{T_f - 1}\sum_{f=2}^{T_f}\sigma\left[m(f,i,j)\right], \quad (1)$$

where the total number of frames is T_f and $\sigma\left[m(f,i,j)\right]$ is the standard deviation of the luminance difference between the present frame, $y(f,i,j)$, and the previous one, $y(f-1,i,j)$. The MAD feature represents the temporal difference between successive frames. The Average MAD $\left(\overline{MAD}\right)$ corresponds to the average of MAD for all frames in the video sequence with $f > 1$, as follows

$$\overline{MAD} = \frac{1}{(T_f - 1)}\sum_{f=2}^{T_f}\frac{1}{MN}\sum_{i=1}^{M}\sum_{j=1}^{N}\left|y(f,i,j) - y(f-1,i,j)\right|. \quad (2)$$

The weighted MAD $(MADp)$ describes the motion of the actual frame (f) relative to the previous frame $(f-1)$ and the average MADp $\left(\overline{MADp}\right)$ over all frames is denoted by

$$\overline{MADp} = \frac{1}{(T_f - 1)}\sum_{f=2}^{T_f}\left(\frac{MAD_f}{MAD_{f-1}}\right),\ f = 2:T_f. \quad (3)$$

Finally, we propose a nonlinear sigmoid-type mathematical model to describe the relationship among the A, B, Z, TI, $\overline{MAD}$, and $\overline{MADp}$ features, which is based on our empirical studies involving the mathematical manipulation of such descriptors. The analytical expression is then given as

$$NRVQA = 1 + \frac{\beta_1}{1+\exp\left(\beta_2 A+\beta_3 B+\beta_4 Z+\beta_5 TI+\beta_6\overline{MAD}+\beta_7\overline{MADp}+\beta_8\right)}, \quad (4)$$

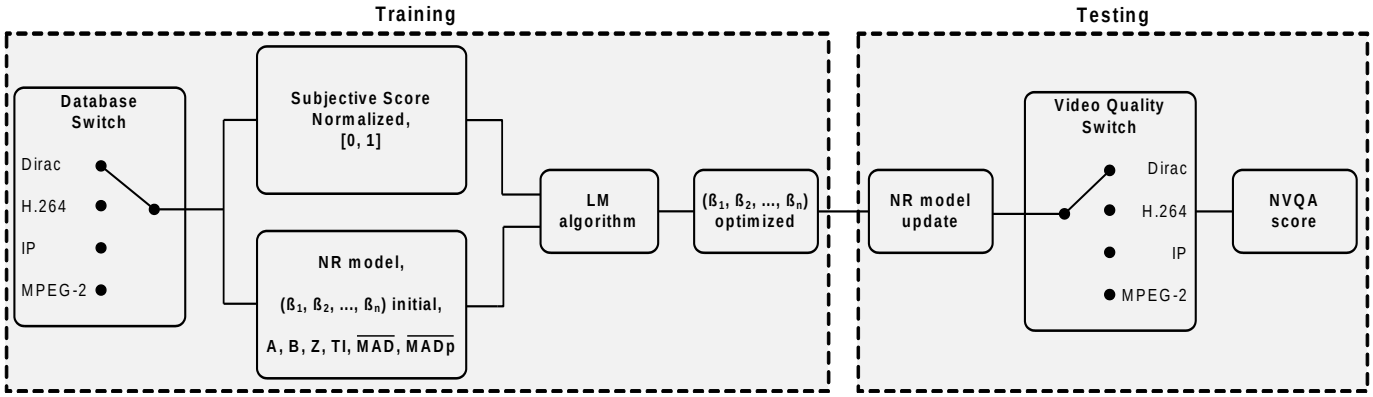


Fig. 1. No-Reference Video Quality Assessment method with database source switched during training phase and video quality score optimized through of the LM algorithm.

where β_1 to β_8 are optimized with the LM method (cf. [6]), used to solve the nonlinear least squares problem.

III. Results and Discussion

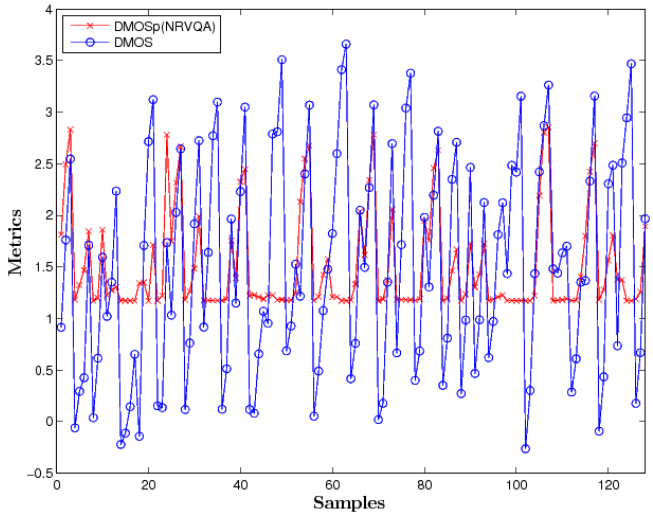
We use the Pearson Linear Correlation Coefficient (PLCC) as the statistical method to measure the performance (accuracy) between our objective metric and the subjective scores (DMOS) of the IVP 1088p@25 database [8]. This database contains 128 videos samples of ten different video sequences in High Definition (HD) and progressive mode with resolution of 1920x1088 pixels and 25 frames per second (PFS). In addition, we choose the IVP 1088p@25 database because it contains four kinds of distortions in HD resolution, such as packet loss of H.264 streams transmitted over IP (28 samples), sequences encoded with the Dirac wavelet (30 samples), H.264 (40 samples) and MPEG-2 (30 samples). The experimental procedure occurs in two steps: a) the calculation of the constants β for each video subset followed by b) the calculation of the PLCC coefficients for performance check. First, the video database was split into five categories, named IP, Dirac, H.264, MPEG-2 and "whole database", this last one representing the combination of all video sequences available in the database. Then, in the training phase, each one of these five categories was further divided in three subsets, named Group 1 (G1), Group 2 (G2) and "T" (Group 1 + Group 2). For instance, both G1 and G2 have 15, 14, 20, and 15 different videos samples, respectively, which contain distorted sequences from the categories Dirac, IP, H.264, and MPEG-2, respectively. The whole database contains the total amount of available video sequences (128 samples).

The quality calibration is performed using the mapping of objective scores into the DMOS scale through the cubic polynomial function, according to the newest VQEG recommendation for HD video content [9].

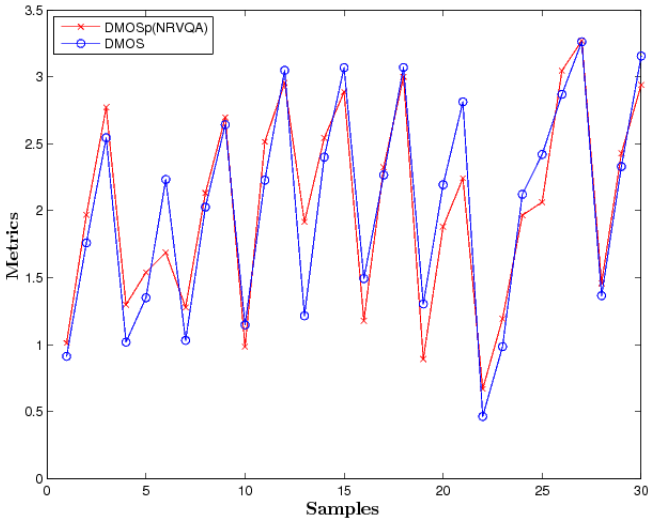
The performance results based on the calculation of PLCC coefficients of our proposed method are compared with the results of other five metrics: Feature-SIMilarity (FSIM), Multi-Scale Structural SIMilarity index (MS-SSIM) [10], Peak Signal-to-Noise Ratio (PSNR), and Structural SIMilarity (SSIM) [11], which are full-reference metrics, and JPEG-NR [7], which is a no-reference metric.

The proposed NRVQA demonstrates the highest performance, in comparison with FR metrics and JPEG-NR for each one of the video subsets, due to the specialized training on a particular category. Such result can be better interpreted by examination of the curves in Fig. 2, where a comparison is made considering the "whole database" and the Dirac subsets, as an example. Fig. 2(a) shows that NRVQA scores do not correlate well when taking all videos in the database. However, it provides a significant improvement when one particular video category is chosen. For instance, for the Dirac category the NRVQA method gives better correlation regarding to DMOS, as seen in Fig. 2(b), with PLCC = 0.936.

Data shown in bold type in Table I give results for the



(a)



(b)

Fig. 2. Comparison between predicted DMOS (NRVQA) and DMOS as training and testing. (a) The whole video database is selected for training and later for testing. (b) Only the Dirac video subset is selected for training and later for testing.

PLCC and point out to the best score in each category and also in each video subset. The results of Table I show that the proposed method presents better accuracy than other methods, when the algorithm is trained for a particular subset, such as Dirac, IP, H.264, or MPEG-2. However, when the metric is applied on the "whole database" subset it presents a lower performance than the FR metrics used for comparison. Yet, its results are still better than JPEG-NR result (e.g., for the G1 subset as training and G2 as testing, JPEG-NR gives a PLCC of 0.287, while our metric delivers 0.464). Thus, when all video sequences ("whole database") are taken into consideration, the FR metrics presents better performance, as can be seen, for instance, from the results obtained in FSIM, where PLCC equals to 0.693, 0.71, and 0.69 for G1, G2, and T, respectively. However, even when a group is used as training and the other used as testing on the Dirac, IP, H.264,

TABLE I. COMPARISON OF ACCURACY (PLCC) BETWEEN FULL-REFERENCE AND NO-REFERENCE METRICS FOR IVP 1088P@25 VIDEO DATABASE [8]

Metric	Train-ing	Testing														
		Dirac			IP			H.264			MPEG-2			Whole database		
		G1[a]	G2[b]	T[c]	G1[a]	G2[b]	T[c]	G1[a]	G2[b]	T[c]	G1[a]	G2[b]	T[c]	G1[a]	G2[b]	T[c]
FSIM		0.919	0.916	0.897	0.539	0.531	0.51	0.846	0.915	0.867	0.894	0.934	0.874	**0.693**	**0.710**	**0.690**
MS-SSIM		0.910	0.881	0.857	0.481	0.514	0.466	0.790	0.884	0.835	0.714	0.814	0.749	0.639	0.674	0.649
PSNR		0.885	0.900	0.868	0.631	0.689	0.63	0.846	0.882	0.864	0.761	0.736	0.714	0.665	0.637	0.648
SSIM		0.921	0.891	0.878	0.449	0.502	0.465	0.829	0.903	0.861	0.709	0.836	0.747	0.646	0.674	0.646
JPEG-NR		0.704	0.607	0.634	0.423	0.406	0.288	0.685	0.608	0.638	0.767	0.720	0.735	0.154	0.287	0.140
Proposed (NRVQA)	G1	**0.942**	0.954	0.930	**0.982**	0.637	0.734	**0.909**	0.890	0.899	**0.962**	0.930	0.935	0.444	0.464	0.451
	G2	0.810	**0.989**	0.889	0.491	**0.935**	0.408	0.824	**0.926**	0.883	0.839	**0.988**	0.905	0.398	0.562	0.471
	T[c]	0.959	0.959	**0.936**	0.831	0.932	**0.898**	0.834	0.914	**0.885**	0.941	0.983	**0.951**	0.433	0.514	0.466

[a]Group 1; [b]Group 2; [c]Total group (G1 + G2).

and MPEG-2 sequences, our method presents the best (G1) or equivalent (G2) accuracy than the other metrics.

IV. CONCLUSION

This work proposes a new no-reference video quality assessment method based on an analytical approach, where the spatial-temporal features are weighted by values obtained in the training phase through the LM algorithm employed to solve the nonlinear least squares problem. The experimental results show that the NRVQA method presents best performance in terms of accuracy (PLCC) in comparison with other full-reference and the JPEG-NR metrics, when the method is applied to specific video categories rather than to a general video database. In principle, this seems to be a drawback of the metric. However, as most of the video content being delivered today is compressed by one specific technique and transmitted over a particular channel, it indeed turns to be an improved approach. At the same time, if one is able to identify the type of compressed stream being used and calculate beforehand the constants β for the corresponding spatial-temporal features of a video database, the proposed switching scheme proposed in Fig. 1 can be applied to most video contents delivered in present days. The comparison between our results and those from other NR metrics in the literature is however difficult due to the use of different mapping functions and different available databases, which are obtained under different conditions and video parameters. In addition, we use the cubic mapping function, as recommended by VQEG, while most works used the logistic mapping function that has been overtaken.

As the proposed method does not require information of the video reference, it is suited for monitoring the video quality at the receiver side. For instance, in digital TV broadcast or mobile systems (where an increasing video content is being transmitted to devices, such as smartphones, tablets and mobile PCs), the video quality scores can be sent back to the central station, via a return channel for further analysis and possible local corrections of the video distortion.

ACKNOWLEDGMENT

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 Image and Video Processing Laboratory from the Chinese University of Hong Kong for using the IVP database.

REFERENCES

[1] H. R. Wu, K. R. Rao, and A. A. Kassim, "Digital Video Image Quality and Perceptual Coding," *Journal of Electronic Imaging*, vol. 16, no. 3, 2007.

[2] F. Yang, S. Wan, Y. Chang, and H. R. Wu, "A novel objective no-reference metric for digital video quality assessment," *Signal Processing Letters, IEEE*, vol. 12, no. 10, pp. 685-688, 2005.

[3] F. Yang, S. Wan, Q. Xie, and H. R. Wu, "No-Reference Quality Assessment for Networked Video via Primary Analysis of Bit Stream," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 20, no. 11, pp. 1544-1554, Nov. 2010.

[4] T. Brandão and M. P. Queluz, "No-Reference Quality Assessment of H.264/AVC Encoded Video," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 20, no. 11, pp. 1437-1447, Nov. 2010.

[5] G. Valenzise, S. Magni, M. Tagliasacchi, and S. Tubaro, "No-Reference Pixel Video Quality Monitoring of Channel-Induced Distortion," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 22, no. 4, pp. 605-618, Apr. 2012.

[6] D. W. Marquardt, "An Algorithm for Least-Squares Estimation of Nonlinear Parameters," *SIAM Journal on Applied Mathematics*, vol. 11, no. 2, p. 431, 1963.

[7] Z. Wang, H. R. Sheikh, and A. C. Bovik, "No-reference perceptual quality assessment of JPEG compressed images," in *Image Processing. 2002. Proceedings. IEEE International Conference on Image Processing*, 2002, vol. 1, p. vol.1 I-477 - I-480.

[8] S. Li and L. Ma, "Full-reference Video Quality Assessment by Decoupling Detail Losses and Additive Impairments," *IEEE Transactions on Circuits and Systems for Video Technology*, no. 99, 2012.

[9] Video Quality Experts Group (VQEG), "Report on the validation of video quality models for high definition video content, version 2.0," 2010.

[10] Z. Wang, E. P. Simoncelli, and A. C. Bovik, "Multiscale structural similarity for image quality assessment," in *The Thrity-Seventh Asilomar Conference on Signals, Systems & Computers, 2003*, 2003, vol. 2, no. 1, pp. 1398-1402.

[11] Z. Wang, A. C. Bovik, H. R. Sheikh, and E. P Simoncelli, "Image quality assessment: from error visibility to structural similarity.," *IEEE Transactions on Image Processing*, vol. 13, no. 4, pp. 600-612, 2004.