A Correlation-Based No-Reference Packet-Loss Metric

Dario D. R. Morais, Alexandre F. Silva and Mylène C. Q. Farias

***Abstract*— In the last few years, there has been a growing interest in automatic methods (metrics) that blindly estimate the quality of digital videos, specially in applications where the videos are digitally transmitted over wired or wireless channels and the original (reference) video signal is not available at the receiver. Currently, there are few methods that are able to identify and estimate the strength of temporal artifacts (e.g. packet-loss and jerkiness) introduced during a digital transmission. In this paper, we present a blind (no-reference) video quality assessment metric that estimates the impact of packet-loss artifacts on quality. The proposed algorithm performs a comparison of inter and intra-block correlations, considering blocks of sizes 8×8, 16×16, and 32×32. The proposed metric was tested on a database containing videos with packet-loss artifacts with different strengths and time durations. Results show that the proposed metric is able to estimate the impact that packet-loss artifacts have on the overall perceived quality, outperforming several metrics available in the literature.**

***Keywords*— video quality metrics, artifacts, quality assessment, NR quality metrics, packet-loss.**

I. Introduction

The most accurate way to determine video quality is by measuring it using psychophysical experiments with human subjects. Unfortunately, these experiments are expensive, time-consuming and hard to incorporate into an automatic system. Therefore, there is currently a need for fast and accurate algorithms (objective video quality metrics) that can provide a measure of the video quality, as perceived by human users. As expected, quality metrics have an important role in communications quality control systems. It is worth pointing out that the quality of the received content is one of the most important factors that determines the user's satisfaction and, consequently, to the acceptability of the service [1].

Generally, objective quality metrics can be divided in three different categories, according to the availability of the original (reference) video signal: Full Reference (FR) methods, Reduced Reference (RR) methods, and No-Reference (NR) methods. In the FR approach, the reference video is available at the measurement point (receiver) and the difference between reference and test video can be used to estimate quality. In the RR approach, only part of the reference is available through an auxiliary channel. Finally, in the NR approach, the reference video is not available and the quality estimate is obtained exclusively from the received video [2].

For several years, the networking community has been quantifying transmission errors using simple metrics, such as bit error rate (BER) or packet-loss rate (PLR). Likewise, until recent years, quality measurements in the signal processing community had been limited to a few objective measures, such as peak signal-to-noise ratio (PSNR), mean squared error (MSE), and total squared error (TSE), supplemented by limited subjective evaluation. Unfortunately, although these metrics are relevant for data links and generic signals, in which every bit is equally important, they are not considered good estimates of the user's opinion about the received multimedia content [3]. One of the major reasons why these metrics do not perform as desired is because they do not incorporate human visual system (HVS) features in their computation. Measurements produced by these metrics are simply based on a pixel to pixel (or bit to bit) comparison of the data, without considering what is the content and the relationships among the pixels in an image (or frames). They also do not consider how spatial and frequency artifact characteristics are perceived by human observers [3].

Video quality is still far from being a mature research topic and limited success has been reported for sophisticated models tested under strict conditions, which include a limited range of distortions and video material. A common approach used by many NR metrics consists of estimating the strength of the most relevant artifacts (e.g, blockiness, blurriness, noise, and ringing) [4] and, then, combining them to obtain an estimate of the overall video quality. Nevertheless, very few metrics deal with artifacts introduced by digital transmission, like packet-loss and jerkiness [5]. Babu et al. [6] studied the effect of blockiness (block-edge) and packet-loss artifacts in video streaming applications. Kanumuri et al. [7] proposed an RR method that models how multiple packet losses affect video quality of a H.264 codec video bitstream.

Leszczuk et al. [8] studied how to assess the quality of high definition videos in a streaming scenario. They concluded that the perceived video quality is prone, not only to packet losses, but also to the temporal and spatial patterns of these artifacts. Staelens et al. [9] proposed a bitstream-based NR method that uses a genetic programming-based symbolic regression. They reported that the most relevant parameters to video quality are: duration of the distortions, percentage of lost picture slices lost, type of frame (I, P or B) that suffered the loss, number of picture slices, number of B-frames between the I-frames, and number of consecutive slice drops.

Liu et al. [10] proposed an FR method that estimates the perceptual distortion caused to packet loss and error propagation in each individual frame. Their method is based

D.D.R. Morais and M.C.Q. Farias are with the Department of Electrical Engineering of the University of Brasilia, while A.F. Silva is with the Department of Computer Science of the University of Brasilia. E-mails: ddrmorais@gmail.com, mylene@ene.unb.br, alexandrefieno@gmail.com.

on a just noticeable difference (JND) model that measures the impact of coding artifacts and error propagation on perceptual video quality. Rui et al. [11] proposed two NR methods that estimate the strength of packet-loss artifacts in videos, taking into account the spatial and temporal discontinuities caused by this artifact.

Although there are several packet-loss metrics available in the literature, their accuracy performance is still not satisfactory, specially for the NR scenario. In this paper, we propose a robust no-reference video quality metric that estimates the strength of 'packet-loss' artifacts using a correlation-based approach. More specifically, the work is based on a previous *image* blockiness metric that estimates the strength of blockiness artifacts by comparing the inter and intra-block correlations. Instead of only considering 8×8 block sizes, the proposed metric uses three types of blocks (8×8, 16×16, 32×32). Also, a non linear SVR regression model is used to combine the contributions of each block size in order to obtain a quality estimate that has a good correlation with subjective scores provided by human subjects.

The paper is divided as follows. In Section II, we describe the previous blockiness metric. Section III describes the adaptation of the metric to packet-loss. Section IV presents the simulations and discusses the results. Finally, Section V details our conclusions.

II. Blockiness Metric

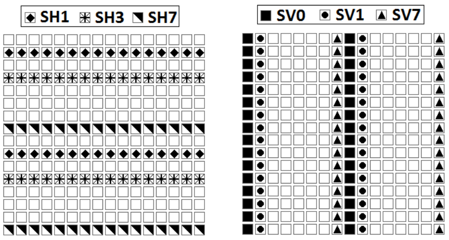
Vlachos [12] proposed an NR metric that estimated the strength of blockiness artifacts by comparing the cross-correlation of pixels inside (intra) and outside (inter) the borders of the coding blocking structure of a frame. In Vlachos'algorithm, the frame $Y(i,j)$ was partitioned into 8×8 blocks and simultaneously sampled in vertical and horizontal directions. This sampling structure assumed that all visible blockiness artifacts had a visible border, what was not always the case.

In a previous work Farias [4], [13] modified Vlachos' algorithm making it possible to take into account cases in which only one of the borders of the blocking structure was visible. Instead of down-sampling the frame simultaneously, the algorithm proposed by Farias split the process into separate vertical and horizontal downsampling processes. As a consequence, the frame was downsampled separately in the vertical and horizontal directions, generating a vertical downsampled image (SV) and a horizontal downsampled image (SH). The two downsampled *images* were computed using the following equations:

$$SH_m = \{Y(i,j) : m = i \mod 8\}. \quad (1)$$

$$SV_n = \{Y(i,j) : n = j \mod 8\}. \quad (2)$$

where (i,j) are the horizontal and vertical co-ordinates and mod is the module operation. This way, SV_n and SH_m are images that contain a subset of pixels with coordinates congruent to 8, either horizontally or vertically respectively. The subscripts m and n can be viewed as the corresponding horizontal and vertical phases, respectively.



(a) (b)

Fig. 1

Frame downsampling structure for: (a) horizontal and (b) vertical directions.

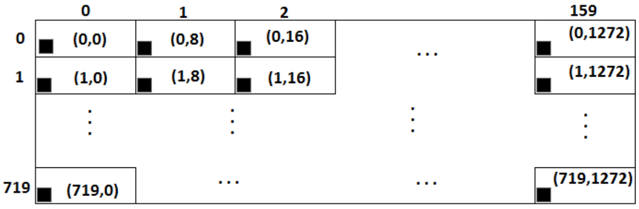


Fig. 2

Illustration of vertical downsampling process used to obtain the sub-image SV_0.

Figures 1 (a) and (b) display the sampling structures used by Farias' algorithm for the horizontal (SH_m) and vertical (SV_n) directions, respectively. The image shows a 16×16 area of the frame, containing four 8×8 blocks. Six sub-images are generated by downsampling pixels located at the positions indicated by the six different symbols. Therefore, different symbols generate different sub-images. The set of inter-block pixels in the vertical direction corresponds to the sub-images SV_0 and SV_7 (Fig. 1 (b)), while the set of inter-block pixels in the horizontal direction corresponds to the sub-images SH_1 and SH_7 (Fig. 1(a)). The set of intra-block pixels in the vertical direction corresponds to the sub-images SV_0 and SV_1 (Fig. 1 (b)), while the set of intra-block pixels in the horizontal direction corresponds to the sub-images SH_1 and SH_3 (Fig. 1 (a)).

Given that interlaced videos were used by Farias, the symbols in the horizontal downsampling structure (see Fig. 1 (a)) are 2 pixels apart, instead of only one pixel like in the vertical downsampling structure (see Fig. 1 (b)). For progressive videos, the symbols should be one pixel apart for both directions. Fig. 2 displays how the sub-image SV_0 is obtained. In this example, the original frame has 1280×720 pixels and the vertically-downsampled sub-image has 160×720 pixels.

The correlation between two images, I_1 and I_2, was given by the following expression:

$$C_{I_1,I_2}(i,j) = F^{-1}\left(\frac{F^*(I_1(i,j)) \cdot F(I_2(i,j))}{|F^*(I_1(i,j))F(I_2(i,j))|}\right), \quad (3)$$

where F and F^{-1} denote the forward and inverse two dimensional discrete Fourier transform, respectively, and $*$ denotes the complex conjugate. The magnitude of the highest peak is a measure of the correlation between I_1 and I_2. But, before the maximum is calculated, the array elements is filtered using a Hamming window, what forces the elements to a constant value around the borders.

To estimate the blockiness signal strength, Farias measured the correlation between the intra- and inter-block sub-images in both directions. For the vertical direction, the correlation was calculated using the following equations:

$$PV_{intra} = \max_{i,j} \{C_{SV_0,SV_1}(i,j)\}, \tag{4}$$

$$PV_{inter} = \max_{i,j} \{C_{SV_7,SV_0}(i,j)\}. \tag{5}$$

The horizontal correlations, PH_{inter} and PH_{intra}, were obtained in a similar way:

$$PH_{intra} = \max_{i,j} \{C_{SH_1,SH_3}(i,j)\}, \tag{6}$$

$$PH_{inter} = \max_{i,j} \{C_{SH_7,SH_1}(i,j)\}. \tag{7}$$

Then, the blockiness measure for one frame was given by:

$$S_{bloc} = \frac{PV_{intra} + PH_{intra}}{PV_{inter} + PH_{inter}}. \tag{8}$$

For frames with no blockiness, the value of PV_{intra} was close to PV_{inter} and PH_{intra} was close to PH_{inter}. As blockiness was introduced, the values of PV_{inter} and PH_{inter} became smaller and, consequently, the value of the blockiness metric increased.

Finally, the blockiness measure for the set of all frames was obtained by averaging the measures over all frames:

$$\hat{S}_{bloc} = \frac{1}{NF} \sum_{nf=0}^{NF} S_{bloc}(nf), \tag{9}$$

where nf refers to the frame number and NF is the total number of frames.

III. Proposed packet-loss metric

The proposed no-reference packet-loss metric is based on the blockiness metric described in the previous section. To adapt the metric proposed by Farias [4], [13] to measure packet-loss (instead of blockiness), we first vary the size of the downsampling structure. Since videos compressed with modern codecs (like H.264 and H.265) use macroblocks of several sizes, we generalize the algorithm proposed by Farias for 8×8, 16×16, and 32×32 block sizes. Figures 3 (a) and (b) show the 8×8 vertical and horizontal downsampling frame structures. Again, the dark symbols in the grids correspond to pixels in the resulting downsampled sub-images. The sampling structures for 16×16 and 32×32 are similar. Notice that, differently from the algorithm proposed by Farias (see Fig. 1), the proposed algorithm simultaneously downsamples the original frame in both directions, reducing the size of the original image in both dimensions.

A total of 6 downsampled images are obtained after the downsampling process, with three sub-images being obtained from the vertical downsampling (DV_7, DV_0, DV_1) and three sub-images from the horizontal downsampling (DH_7, DH_0, and DH_1). Then, similarly to what was done in the previous section, we calculate the cross-correlation between two sub-images to obtain the blockiness measure for a single frame. More specifically, for the vertical direction, we obtain the inter-block correlation by calculating the correlation between sub-images DV_7 and DV_0 and the intra-block correlation calculating the correlation between sub-images DV_0 and DV_1:

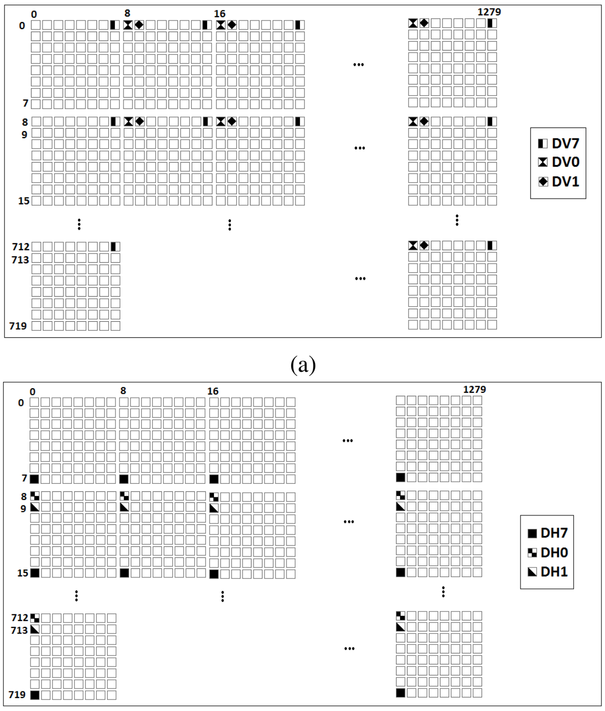


Fig. 3

Frame downsampling structure for the proposed packet-loss metric: (a) vertical and (b) horizontal.

$$PV_{intra,8} = \max_{i,j} \{C_{DV_0,DV_1}(i,j)\}, \tag{10}$$

$$PV_{inter,8} = \max_{i,j} \{C_{DV_7,DV_0}(i,j)\}. \tag{11}$$

Similarly, for the horizontal direction, we obtain the inter-block correlation calculating the correlation between sub-images DH_7 and DH_0 and the intra-block correlation calculating the correlation between sub-images DH_0 and DH_1:

$$PH_{intra,8} = \max_{i,j} \{C_{DH_0,DH_1}(i,j)\}, \tag{12}$$

$$PH_{inter,8} = \max_{i,j} \{C_{DH_7,DH_0}(i,j)\}. \tag{13}$$

The 8×8 block measure for one frame is given by:

$$S_8 = \frac{PV_{intra,8} + PV_{inter,8}}{PH_{intra,8} + PH_{inter,8}} \tag{14}$$

Notice that, given that we are assuming the frames are in a progressive format, there is no shift between the pixels. To obtain a measure for the complete video, we average S_8 for all frames, obtaining $\hat{S}_8$.

Next, we use the same algorithm on blocks of size 16×16 ($\hat{S}_{16}$) and 32×32 ($\hat{S}_{32}$). The final packet-loss metric value ($\hat{S}_{pck}$) is a composition of the measures for the three block sizes ($\hat{S}_8$, $\hat{S}_{16}$, and $\hat{S}_{32}$), which is obtained using a support vector regression (SVR) technique (The SVM function in the R software was used in this work). SVR is a black-box approach, in which the model is not defined upfront but

Fig. 4

SAMPLE FRAMES OF ORIGINAL VIDEOS USED IN THE EXPERIMENT, TOP TO BOTTOM, LEFT TO RIGHT: PARK JOY, INTO TREE, PARK RUN, ROMEO AND JULIET, CACTUS, BASKETBALL, AND BARBECUE.

learned directly from the data (i.e. from the video dataset). We choose to use SVR because similar machine learning-based approaches have been used with success to model complex non-linear perceptual processes related to artifact annoyance [14].

IV. SIMULATIONS AND RESULTS

We have tested the proposed algorithm using data collected in an psychophysical (subjective) experiment performed earlier [15]. The database contains seven high definition original videos (see Fig. 4), with spatial resolution of 1280×720 and temporal resolution of 50 frames per second (fps). The videos are all ten seconds long and were chosen with the goal of generating a diverse video content, as recommended in the "Final Report of VQEG on the validation of objective models multimedia quality assessment (Phase I)" [16]. The test videos consist of versions of the original videos degraded with packet-loss artifacts at different packet loss percentages (0.7%, 2.6%, 4.3%, and 8.1%) and with different durations (M = 4, 8, and 12 frames). Subjects in this experiment were asked to rate the annoyance of artifacts in the test videos, using a scale from '0' to '100'. The scores provided by subjects are averaged for each test video, providing mean annoyance values (MAVs).

To train the SVR, we use a k-fold cross validation setup. We split the dataset in k equally sized non-overlapping sets. We then run the training k times. In each time, a different fold is used as test set and the remaining $k-1$ folds are used for training. This way, each data point has a chance of being validated against the other [17]. In our tests, we set k to 10, thereby running 10 repetitions of the training. We then compute the correlation between the subjective data (MAVs) and the model predictions for each run and calculate their average, which is used as the model performance measure. The SVR has trained on MAVs and it has returned Pearson correlation coefficient (PCC) and Spearman correlation coefficient (SCC) values equal to 0.825 and 0.778, respectively.

For comparison purposes [16], we select the following quality metrics:

- 3 FR metrics: peak signal-to-noise ratio (PSNR), mean squared error (MSE) and multi-scale structural similarity (MS-SSIM) [18];
- 2 NR packet-loss metrics: Rui et al. [11] and Babu et al. [6].

We choose PSNR and the MSE because they are very popular fidelity metrics in the image and video processing community [16], [19]. MS-SSIM is a popular extension of the SSIM paradigm that has been used with success to estimate video quality [18]. The metrics by Rui et al. [11] and Babu et al. [6] were chosen because they are both packet-loss nr metrics.

To evaluate the performance of the tested metrics, in Table I we report the PCC, the SCC, and the Root mean squared error (RMSE) values [16] computed for the scores predicted with the tested quality metrics and the MAVs from the earlier experiment [15]. Results show that the proposed metric has a much better performance than the other metrics.

Figure 5 shows graphs of the outputs of the individual quality metrics versus the MAVs from the earlier experiment. The graphs corresponding to MSE and PSNR (Fig. 5 (a) and (b)) show a high concentration of points at the center and left regions, indicating that both PSNR and MSE are poor annoyance predictors for packet-loss artifacts. Surprisingly, the graph for the FR metric MS-SSIM and the packet-loss metrics by Rui and Babu also show a large spread of points. In particular, the graph for the metric by Babu (Fig. 5 (d)) shows vertical lines that indicate that test videos with different MAVs obtain a very similar packet-loss score. On the other hand, the graph of the proposed metric (Fig. 5 (f)) shows a much better performance.

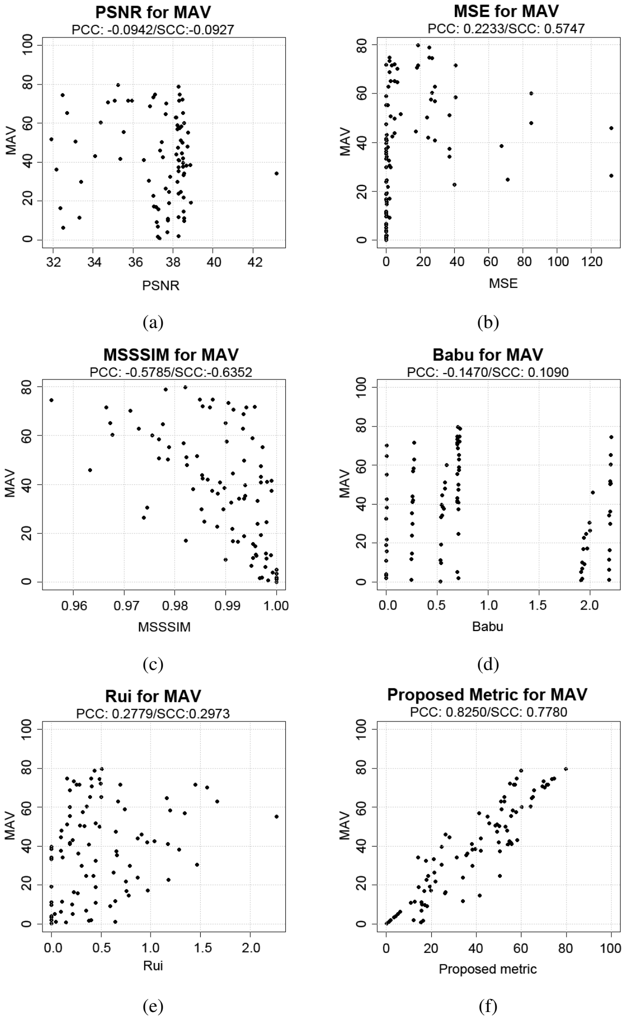
TABLE I

PEARSON (PCC) AND SPEARMAN (SCC) CORRELATION COEFFICIENTS, AND RMSE FOR ALL TESTED METRICS.

Metrics	**PCC**	**SCC**	**RMSE**
PSNR	-0.0942	-0.0927	21.9297
MSE	0.2233	0.5747	39.3895
MS-SSIM	-0.5785	-0.6352	43.8234
Babu	0.1470	-0.1090	43.9555
Rui	0.2779	0.2973	44.1735
Proposed Metric	0.8250	0.7780	13.5173

V. CONCLUSIONS AND FUTURE WORK

In this paper, we have proposed a no-reference packet-loss video quality metric. The proposed metric is a modification of a correlation-based blockiness metric, proposed by Farias, that estimates blockiness by comparing the cross-correlation of pixels inside (intra) and outside (inter) the borders of the 8×8 coding blocking structure. Unlike the blockiness metric, the downsampling structure of the proposed packet-loss metric uses 3 block sizes (8×8, 16×16, and 32×32). The output of the packet-loss metric is a composition of the measures for the three block sizes which is obtained using an SVR technique. Results show that the proposed metric has a better accuracy performance than the other tested metrics, with correlation coefficients above 0.825. Future work includes the combination of this metric with other metrics that estimate strengths of different artifacts, such as jitter, blockiness, blurriness, and ringing.



(a)

(b)

(c)

(d)

(e)

(f)

Fig. 5

QUALITY METRIC RESULTS FOR ALL TEST SEQUENCES: (A) PSNR, (B) BABU, (C) MSE, (D) RUI, (E) MS-SSIM, AND (F) PROPOSED METRIC.

ACKNOWLEDGMENTS

This work is supported in part by Conselho Nacional de Desenvolvimento Científico e Tecnológico (CNPq), in part by Coordenação de Aperfeiçoamento de Pessoal de Nível Superior (CAPES), and in part by the University of Brasília (UnB).

REFERENCES

[1] Weisi Lin and C-C Jay Kuo, “Perceptual visual quality metrics: A survey,” *Journal of Visual Communication and Image Representation*, vol. 22, no. 4, 2011.

[2] Mylene CQ Farias and Sanjit K Mitra, “Perceptual contributions of blocky, blurry, noisy, and ringing synthetic artifacts to overall annoyance,” *Journal of Electronic Imaging*, vol. 21, no. 4, pp. 043013–043013, 2012.

[3] Zhou Wang and Alan C Bovik, “Mean squared error: love it or leave it? a new look at signal fidelity measures,” *Signal Processing Magazine, IEEE*, vol. 26, no. 1, pp. 98–117, 2009.

[4] Mylene CQ Farias and Sanjit K Mitra, “No-reference video quality metric based on artifact measurements,” in *Image Processing, 2005. ICIP 2005. IEEE International Conference on*. IEEE, 2005, vol. 3, pp. III–141.

[5] R Venkatesh Babu, Andrew Perkis, and Odd Inge Hillestad, “Evaluation and monitoring of video quality for uma enabled video streaming systems,” *Multimedia Tools and Applications*, vol. 37, no. 2, pp. 211–231, 2008.

[6] R Venkatesh Babu, Ajit S Bopardikar, Andrew Perkis, and Odd Inge Hillestad, “No-reference metrics for video streaming applications,” in *International Workshop on Packet Video*, 2004.

[7] Sandeep Kanumuri, Sitaraman G Subramanian, Pamela C Cosman, and Amy R Reibman, “Predicting h. 264 packet loss visibility using a generalized linear model,” in *Image Processing, 2006 IEEE International Conference on*. IEEE, 2006, pp. 2245–2248.

[8] Mikołaj Leszczuk, Lucjan Janowski, Piotr Romaniak, and Zdzisław Papir, “Assessing quality of experience for high definition video streaming under diverse packet loss patterns,” *Signal Processing: Image Communication*, vol. 28, no. 8, pp. 903–916, 2013.

[9] Nicolas Staelens, Dirk Deschrijver, Ekaterina Vladislavleva, Ben Vermeulen, Tom Dhaene, and Piet Demeester, “Constructing a no-reference h. 264/avc bitstream-based video quality metric using genetic programming-based symbolic regression,” *Circuits and Systems for Video Technology, IEEE Transactions on*, vol. 23, no. 8, pp. 1322–1333, 2013.

[10] Tao Liu, Hua Yang, Alan Stein, and Yao Wang, “Perceptual quality measurement of video frames affected by both packet losses and coding artifacts,” in *Quality of Multimedia Experience, 2009. QoMEx 2009. International Workshop on*. IEEE, 2009, pp. 210–215.

[11] Hua-xia Rui, Chong-rong Li, and Sheng-ke Qiu, “Evaluation of packet loss impairment on streaming video,” *Journal of Zhejiang University SCIENCE A*, vol. 7, no. 1, pp. 131–136, 2006.

[12] T Vlachos, “Detection of blocking artifacts in compressed video,” *Electronics Letters*, vol. 36, no. 13, pp. 1106–1108, 2000.

[13] Hugo Tadashi M Kussaba and Mylene CQ Farias, “Blind estimation of blocking artifacts in digital videos,” *Latin Display 2010*, 2010.

[14] Paolo Gastaldo, Rodolfo Zunino, and Judith Redi, “Supporting visual quality assessment with machine learning,” *EURASIP Journal on Image and Video Processing*, vol. 2013, no. 1, pp. 1–15, 2013.

[15] Mylène CQ Farias, I Heynderickx, BL Macchiavello Espinoza, and JA Redi, “Visual artifacts interference understanding and modeling (varium),” in *Seventh international workshop on video processing and quality metrics for consumer electronics*, 2013, vol. 1.

[16] Video Quality Experts Group, “Final report from the video quality experts group on the validation of objective models of video quality assessment - Phase II,” Tech. Rep., http://ftp.crc.ca/test/pub/crc/vqeg/, 2003.

[17] Payam Refaeilzadeh, Lei Tang, and Huan Liu, “Cross-validation,” in *Encyclopedia of database systems*, pp. 532–538. Springer, 2009.

[18] Zhou Wang, Eero P Simoncelli, and Alan C Bovik, “Multiscale structural similarity for image quality assessment,” in *Signals, Systems and Computers, 2004. Conference Record of the Thirty-Seventh Asilomar Conference on*. Ieee, 2003, vol. 2, pp. 1398–1402.

[19] Stefan Winkler and Praveen Mohandas, “The evolution of video quality measurement: from psnr to hybrid metrics,” *Broadcasting, IEEE Transactions on*, vol. 54, no. 3, pp. 660–668, 2008.

[20] Alexandre F. Silva, Mylene C. Q. Farias, and Judith A. Redi, “Assessing the influence of combinations of blockiness, blurriness, and packet loss impairments on visual attention deployment,” in *IS&T/SPIE Electronic Imaging*, 2015, vol. 9394, pp. 93940Z–93940Z–11.