Journal of Electronic Imaging

JElectronicImaging.org

Reduced-reference video quality assessment using a static video pattern

Michail-Alexandros Kourtis
Harilaos Koumaras
Fidel Liberal

SPIE. IS&T imaging.org

Michail-Alexandros Kourtis, Harilaos Koumaras, Fidel Liberal, "Reduced-reference video quality assessment using a static video pattern," *J. Electron. Imaging* **25**(4), 043011 (2016), doi: 10.1117/1.JEI.25.4.043011.

Reduced-reference video quality assessment using a static video pattern

Michail-Alexandros Kourtis,[a,*] Harilaos Koumaras,[a] and Fidel Liberal[b]
[a]National Center for Scientific Research "Demokritos," Institute of Informatics and Telecommunications, Agia Paraskevi, Patriarxou Grigoriou, Athens 15310, Greece
[b]University of the Basque Country, Department of Communications Engineering, Alameda de Urquijo, Bilbao 48013, Spain

Abstract. A reduced-reference video quality assessment (VQA) method was proposed by using structural similarity (SSIM) index as a tool to extract features from both the original and the target video sequences, using a reference video pattern. The method is suitable for monitoring the video quality in real time and across the service provision chain. The performance of the proposed method was evaluated using a large experimental set of reference and nonreference video sequences and achieves an accuracy higher than 2.56% in comparison to SSIM. Additionally, comparison to subjectively evaluated scores of Laboratory for Image and Video Engineering video quality dataset, based on difference mean opinion scores, shows that the performance of the proposed method is within the range of the full reference VQA methods. [DOI: 10.1117/1.JEI.25.4.043011]

Keywords: video quality assessment; structural similarity; reduced reference; feature extraction; video coding; transcoding.

Paper 15756 received Oct. 9, 2015; accepted for publication Jun. 30, 2016; published online Jul. 19, 2016.

1 Introduction

The thriving interest of consumers for video content has brought the world closer to the universe of digital video provision than ever before. The recent market success of media services such as internet protocol television and video on demand through consumer electronic products, has created a considerable increase of the network traffic, which has caused the network operators to apply restriction policies, creating a controversial issue on the network neutrality and traffic differentiation.

This situation has created the need for more advanced encoding techniques (e.g., H.265/HEVC) and adaptation schemes[1] which achieve higher compression ratios and provide agility to the adaptation of the media service,[2] alleviating the network operators from the media-related network traffic. Thus, during the service provision, the video stream may be needed to be dynamically transcoded at different formats/profiles (e.g., such as in the paradigm of mobile-edge computing), resulting in adapted media services that dynamically fit the current network conditions and the terminal device specifications. However, this in-service transcoding process, which today can be supported by the emerging software-defined networking and network function virtualization techniques,[3] introduces to the media service a wide variety of encoding impairments that degrade the deduced quality level of the encoded media service. Thus, the quality degradation of the media service, caused by the in-service transcoding process, creates the need for defining flexible video quality assessment (VQA) methods that will be able to evaluate the quality with the same accuracy as the well claimed video quality metrics [e.g., structural similarity (SSIM) index[4]]. These VQA methods would be able to assess the quality not only during the initial coding process of the source signal, but also across the media service delivery path to the end user, providing useful feedback both to the content provider for service adaptation actions and to the network operator for optimal traffic steering decisions.

To monitor the quality in real time and across the service provision chain,[5] it is necessary to use flexible VQA tools, which are suitable for in-service integration, evaluating the media service along its network delivery path.

Currently, the available VQA methods are divided into two categories: the subjective and the objective ones. More specifically, the subjective VQA methods[6] are based upon the opinion score of a group of viewers regarding the visual degradation of an encoded video sequence compared to the original uncompressed sequence, establishing them as the primary choice for video quality evaluation tests in terms of reliability, but also practically from a commercial perspective. The subjective video quality evaluation methods are expensive and time-consuming mainly due to their demanding setup within a controlled room/environment with sophisticated apparatus, which leads to the fact that they cannot be commercially exploited, especially within the service provision chain for monitoring purposes of the delivered video service.

Correspondingly, the objective VQA techniques are mathematical computational models that utilize various image characteristics (e.g., luma and chroma),[7–16] or other image statistics (e.g., blockiness),[17] to approximate as well as possible the subjective test results in an efficient and cost-effective way. The objective methods are categorized into three groups determined by their approach and the metric used for the quality assessment: the full reference (FR) ones, the reduced reference (RR) ones, and the no-reference (NR) ones.

The FR methods evaluate the video quality by comparing the frames of the original video and the target video. The

*Address all correspondence to: Michail-Alexandros Kourtis, E-mail: akis.kourtis@iit.demokritos.gr

methods perform multiple channel decomposition of the video signal, where the objective method is applied on each channel, which feature a different weight factor according to the characteristics of the human visual system (HVS), using contrast sensitivity functions, channel decomposition, error normalization, weighting, and finally Minkowski error pooling for combining the error measurements into a single perceived quality estimation.[18] Also, in the bibliography, FR methods for a single channel have been proposed where the proposed objective metric is applied on the video signal, without considering varying weight functions. Some FR metrics that are based on the video structural distortion have been proposed,[19] among which is the widely known SSIM index, which has a very wide range of applicability across many different fields.[20–22] All FR methods, including SSIM, provide higher accuracy and credibility in comparison to the rest of the categories (RR and NR), but in the evaluation process, they require both the original and the encoded video sequences at the same site, making them inappropriate for integration in the service provision chain where the original signal is not available at the end-user site.

The RR methods are able to evaluate the video quality level based on metrics which use only some extracted structural features from the original signal.[23,24] The concept of the RR metrics was introduced by Refs. 7, 24, and 25, where the RR metric was based upon the extraction of various spatial and temporal features of the reference video, which are easily exposed to distortions added by the standard video compression process. RR metrics can be roughly categorized into three categories.

The first category includes all methods based upon models of low-level statistical properties of the original natural image. An RR metric that belongs to this category is described in Ref. 26, which provides a condensed amount of RR information obtained by the comparison of the marginal probability distribution of wavelet coefficients in different wavelet subbands with the probability density function of the wavelet coefficients of the decoded signal, while using the Kullback–Leiber divergence as a distance between distributions.

The second category of RR metrics includes methods that capture visual distortions, to quantify the decoded signal's quality.[27–30] However, this type of metrics only performs well, when there is sufficient knowledge about the degradation process that the signal underwent. It is not efficient to apply these techniques to general cases whose distortion has not been previously assumed.

The third and the last categories of RR metrics are based upon models of the viewer's perception, e.g., the HVS.[31–34] These models exploit and apply different psychological and psychological vision studies on the end users in an attempt to imitate the behavior of subjective test groups.

RR methods are more flexible for in-service integration since they require only partial information of the original video signal, but they have reduced accuracy and credibility in comparison to the FR metrics.

Finally, the NR methods evaluate the video quality on the basis of processing the frames of the target video alone. As they do not require any information from the original video sequence, they can be easily integrated within the service provision chain. However, their performance is limited to specific visual artifacts (e.g., tiling), restricting their range of applicability to special cases only. Thus, from the family of the objective methods, the RR and the NR metrics are more suitable for in-service integration, but they suffer from limited efficiency in comparison to the FRs, which offer high accuracy, but applicability limitations, since they require the original video sequence for assessing the video quality.

It is evident that RR methods require an additional communication channel within the network architecture to transmit the extracted features from the video provider site to the end-user terminals. It is obvious that the required bandwidth depends on the specific RR method. A bandwidth in the range of 1 to 150 kbps is usually required, depending on the RR method and the feature extraction type.[23,35] However, the method described in Ref. 35, which requires under 1-kbps bandwidth, performs very poorly in terms of absolute difference compared to subjective tests.

An alternative technique is that the extracted features (or in general the reference information) will be encapsulated inside the forward link, along with the video transmission. The more features that are extracted and transmitted to the end user, the more accurate is the objective video quality. However, more features require higher bandwidth of the communication network. So in RR methods, there is a trade-off between the accuracy of the VQA and the constraints in the network bandwidth.

In this paper, an RR method is proposed that is suitable for in-service use. The features extracted from both the original and the target video frames are based on the evaluation of the SSIM index. In this respect, the SSIM index is calculated for each frame of the original video using as reference a static white pattern, i.e., a video whose frames are all white. The SSIM index for each frame is transmitted to the end-user site. At the end-user terminal, the SSIM index is calculated between each frame of the received (target) video and the same static white pattern. The proposed RR metric is the ratio of the two SSIM indices. Through experimental measurements, it is shown that the proposed metric has a value very close to the SSIM index as calculated from the comparison of the original and the target video frames. The accuracy of the proposed metric as compared to SSIM depends on the video degradation, i.e., the higher the degradation, the worse is the accuracy. Experimental measurements show that for an acceptable video quality level,[36,37] the mean absolute percentage deviation (MAPD) of the proposed method is lower than 2.56%.

Another advantage of the proposed method is the low bit rate reference information signal that needs to be sent to the end user, which ranges between 400 and 600 bps.

The rest of the paper is organized as follows: Sec. 2 describes the use of SSIM as a feature extraction method from both the original and the target videos and introduces the metric SSIM RR (SRR), which can be directly compared to the original SSIM index. Section 3 presents a qualitative interpretation of the proposed video quality metric SRR. In Sec. 4, the performance evaluation of the proposed method is presented and analyzed, and finally Sec. 5 concludes the paper.

2 Feature Extraction Using Structural Similarity

Among the most reliable objective evaluation metrics is the SSIM, which measures the SSIM between two image

sequences, exploiting the general principle that the main function of the HVS is the extraction of structural information from the viewing field. If x and y are two video signals, then SSIM is defined as

$$\mathrm{SSIM}(x,y)=\frac{(2\mu_x\mu_y+C_1)(2\sigma_{xy}+C_2)}{(\mu_x^2+\mu_y^2+C_1)(c_x^2+\sigma_y^2+C_2)}, \tag{1}$$

where μ_x and μ_y are the mean of x and y, σ_x, σ_y, and σ_{xy} are the variances of x, y and the covariance of x and y, respectively. The constants C_1 and C_2 are defined as

$$C_1=(K_1L)^2C_2=(K_2L)^2, \tag{2}$$

where L is the dynamic pixel range and $K_1=0.01$ and $K_2=0.03$, respectively.

In the typical SSIM index evaluation process, it is assumed that both the original and the target video sequences are available at the same site, as shown in Fig. 1. SSIM(x,y) evaluation for every frame can be based on any software implementation of Eq. (1), where x is the original video sequence ($\mathrm{VS_o}$) in Fig. 1, y is the target video sequence $\mathrm{VS_t}$, and $\mathrm{SSIM_{ot}}$ is their SSIM(x,y) index.

According to Ref. 4, where SSIM index is defined and introduced, SSIM comprises three image characteristics components: the luminance, the contrast, and the structure comparison. Their combination results in the widely used SSIM index. The three separate comparison components are

$$l(x,y)=\frac{2(1+R)}{1+(1+R)^2+\frac{C_1}{\mu_x^2}}, \tag{3}$$

$$c(x,y)=\frac{2\sigma_x\sigma_y+C_2}{\sigma_x^2+\sigma_y^2+C_2}, \tag{4}$$

$$s(x,y)=\frac{\sigma_{xy}+C_3}{\sigma_x\sigma_y+C_3}. \tag{5}$$

The combination of Eqs. (3) luminance, (4) contrast, and (5) structure comparisons results in the SSIM index equation

$$\mathrm{SSIM}(x,y)=[l(x,y)]^a*[c(x,y)]^\beta*[s(x,y)]^\gamma. \tag{6}$$

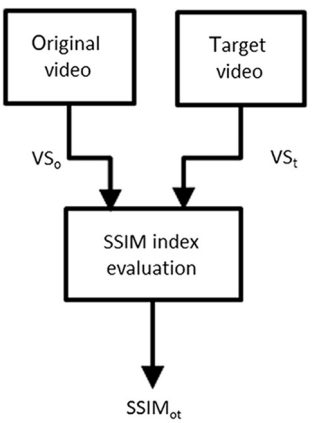

Fig. 1 Typical SSIM index evaluation with original and target video sequences available at the same site.

According to the SSIM index equation, the parameters α, β, and γ adjust the relative importance of each of the three components in the calculation of the SSIM, but for simplicity reasons, the authors of Ref. 4 have selected the case that $\alpha=\beta=\gamma=1$, which results in the well-known expression of the SSIM index. However, the decision that the three components participate equally in the final calculation of SSIM index is not sufficiently justified by the authors in Ref. 4 and it seems that it is a decision reached only for simplicity reasons.

This motivated us to further research the sensitivity analysis of the SSIM accuracy under different α, β, and γ weights. More specifically, considering the purpose of this paper is to develop a flexible metric suitable for in-service applicability, we notice that the parameter γ specifies the importance of the SSIM between signals x and y, which is the most influential factor for the FR requirement in the applicability of the SSIM, since according to the following type, the σ_{xy} factor needs to be calculated for the measurement of Eq. (5).

In the proposed method, in order to investigate the in-service applicability of the SSIM [i.e., in in-service cases where the calculation of Eq. (5) is not feasible due to the lack of the reference signal], we research in this paper the case that $\gamma\to 0$, so the relevant importance of Eq. (5) in the SSIM calculation is limited. Therefore, the requirement for the reference signal to be available together with the encoded signal for the estimation of Eq. (5) ceases to exist, allowing the decomposition of the SSIM index exclusively for the Eqs. (3) and (4) parameters, where no ex parameters exist that require the existence of both the reference and the encoded signal at the same place.

Based on this analysis, in the proposed method (see Fig. 2), the SSIM index is used as a tool to extract features from both the original and the target video sequences using a reference video pattern. In this respect, an initial $\mathrm{SSIM_{or}}$ value is evaluated for every frame at the service provider

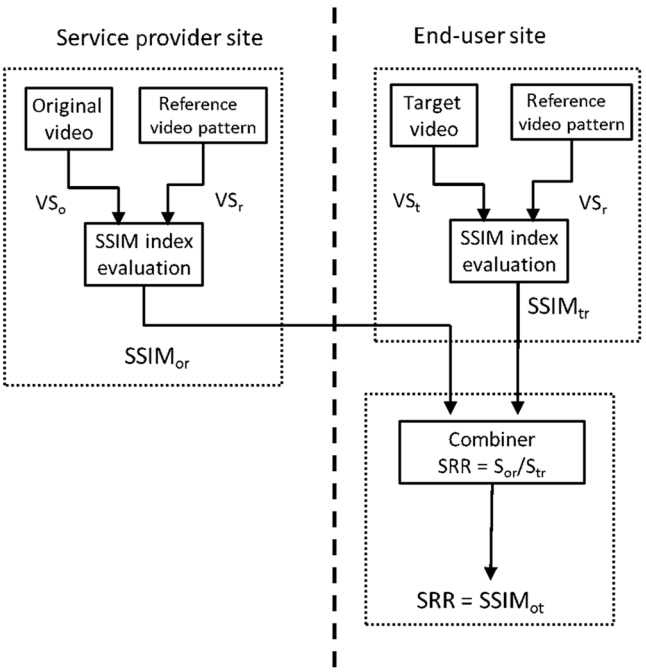

Fig. 2 SRR evaluation method using a reference video pattern as reference at both the service provider and end-user sites.

site, by comparing the VS_o with a video reference pattern (VS_r), e.g., a video sequence of white video frames of the same resolution and frame rate, which is artificially generated [later in this section, the authors test the primary colors (i.e., white, black, red, green, and blue) as reference video patterns and it is shown that the white video pattern performs better than the others]. The evaluation of $SSIM_{or}$ can be based on any software implementation of Eq. (1), where x is VS_o and y is VS_r and refers to each frame. $SSIM_{or}$ can be considered as a feature of each frame of the original video and is sent to the end-user site by any means, i.e., either through the same communication channel as the video, or any other communication channel with a sufficient bandwidth, or by embedding it inside the transport stream of the video sequence. In any case, it is considered that the $SSIM_{or}$ value is recovered at the end-user site.

Referring to Fig. 2, an $SSIM_{tr}$ value is evaluated at the end-user site by comparing the received (target) video signal (VS_t) with a reference video pattern (VS_r), which is identical to the one used at the service provider site and is also artificially generated at the site. The evaluation of $SSIM_{tr}$ is based on the same software implementation of Eq. (1) which used at the transmitter site and refers to each frame of the target video sequence. $SSIM_{tr}$ can be considered as a feature of each frame of the target video.

The ratio of these two SSIM values can be considered as a new metric based on SSIM, namely SRR, i.e.,

$$\mathrm{SRR} = \mathrm{SSIM}_{or}/\mathrm{SSIM}_{tr}. \tag{7}$$

Comparing SRR with the SSIM index between the original and the target video sequences, experimental results in Sec. 5 show that the SRR efficiently approximates the SSIM, with an MAPD <2.56% and satisfactory correlation coefficient values with subjective mean opinion scores (MOS).

Concerning the kind of reference video pattern, the applicability of the SRR method was evaluated based on using the primary colors (i.e., white, black, red, green, and blue) as reference video patterns. The selected primary colors portray a distinct RGB diversity, which is essential to demonstrate and evaluate the behavior of the proposed method, maintaining a simplicity in the implementation and representation of the pattern.

In this framework, the proposed method was applied on the experimental set of the 40 test signals, which are encoded with H-264 at three distinct quantization parameters (QP) values, specifically QP = 12, 22, and 32, each time utilizing a different primary color as a reference video pattern. The QP value regulates how much spatial detail is maintained during the encoding/compression process (modifying, respectively, the quantization step). A low QP value denotes a low quantization step, therefore, almost all the spatial detail of the video is retained, while a high QP value corresponds to high quantization step and the video spatial detail is aggregated resulting in increased distortion and degradation of video quality. For each color and QP value, the proposed SRR method was applied and each time the MAPD value in relevance to the SSIM was calculated.

The experimental results of the process are provided in Table 1, where it is observed that among the primary colors, the white video pattern provides a better performance across different QP values, while the rest of the primary colors perform notably worse than the white. Therefore, considering its better performance than the rest of the colors and its main characteristic as the easiest representation with only one luminance parameter, the white reference video pattern is recommended for implementing the proposed SRR method and it is selected for executing the experimental parts of this paper.

Concerning the channel requirements and the overhead of the proposed method, $SSIM_{or}$ is a number <1 and it can be represented by 2 bytes per frame for an accuracy of four decimal places (10^{-4}). In this case, the required bit rate to be transmitted to the end-user site is 400 bps per 25 frames/s. For an increased accuracy per frame of six decimal places (10^{-6}), 3 bytes are required, resulting in 600 bps for the $SSIM_{or}$. Even 600 bps is significantly lower than the value of ~1 to 150 kbps required for other RR methods, as mentioned in Sec. 1.

3 Qualitative Interpretation of the Proposed Video Quality Metric

In order to interpret the proposed quality metric and its relation to the SSIM index, we consider a video sequence which is encoded at various QP values, resulting in different bit rates and quality levels. The SSIM index between the original and the target video sequence for each frame is denoted as $SSIM_{ot}(QP)$, which is a descending function of QP. Given that the maximum value of $SSIM_{ot}$ is equal to 1, the SSIM variation is an exponential function as shown in Ref. 38. Therefore

$$\mathrm{SSIM}_{ot}(\mathrm{QP}) = e^{-\alpha * \mathrm{QP}}, \tag{8}$$

where α is a coefficient that depends on the content of the video signal.[38]

Table 1 MAPD for 40 test signals with different reference video patterns.

Reference video pattern	R	G	B	QP:12	QP:22	QP:32
Black	0x00	0x00	0x00	0.011049	0.0303	0.099677
White	0xFF	0xFF	0xFF	0.007049	0.010328	0.032666
Red	0xFF	0x00	0x00	0.008156	0.013966	0.042319
Green	0x00	0xFF	0x00	0.009114	0.016522	0.047510
Blue	0x00	0x00	0xFF	0.006726	0.014751	0.044131

The plot of $SSIM_{ot}(QP)$ versus QP is shown in Fig. 3, curve a. As QP increases, more and more information from the original video frame is lost, i.e., the spatial information (i.e., color and intensity of each pixel) of VS_t is lost. In the extreme case, where the information is completely lost, all pixels are equal in color and density, i.e., the VS_t is degraded to a sequence of uniform frames, as for example a white video pattern (similar to the used reference video pattern). In this extreme case, the lowest value of $SSIM_{ot}(QP)$ is the SSIM index between the original video sequence and the reference video pattern, which can be denoted as $SSIM_{or}$.

The previous analysis refers to the comparison between the original and the target video frames. If we consider the comparison between the target video frames and the reference video (i.e., the white video pattern explained earlier), the $SSIM_{tr}(QP)$ index versus QP will be an ascending function.

This is because at low QPs the target frame is almost identical to the original and, therefore, will greatly differ from the (white) reference video pattern, resulting in a low $SSIM_{tr}$ value. Furthermore, as QP increases, the target frame will gradually become similar to the reference white video pattern, resulting in a higher value of $SSIM_{tr}$. For very high QP, the target frame is identical to the reference one and the maximum $SSIM_{tr}$ is equal to 1.

Considering an exponential variation of $SSIM_{tr}$ (similar to $SSIM_{ot}$), it is deduced that

$$SSIM_{tr}(QP) = SSIM_{or} * e^{\alpha * QP}. \tag{9}$$

The plot of $SSIM_{ot}(QP)$ versus QP is shown in Fig. 3, curve b.

From Eqs. (8) and (9), it can be deduced that

$$SSIM_{ot}(QP) * SSIM_{tr}(QP) = SSIM_{or} \rightarrow \\ SSIM_{ot} = SSIM_{or}/SSIM_{tr}. \tag{10}$$

Comparing Eqs. (7) and (10), it is evident that

$$SRR = SSIM_{ot}. \tag{11}$$

That is, in the ideal case, the proposed SRR is equal to the SSIM index. However, in real conditions, the $SSIM_{tr}$ variation differs from Eq. (9), because for high QP values the degraded frame will not become identical to the white reference one. This is due to the fact that the maximum value of $SSIM_{tr}$ in real conditions will not reach the ideal value of 1, as is shown in the plot of $SSIM_{tr}$ in real conditions in curve c of Fig. 3 (dotted line). This deviation will affect the accuracy of Eq. (11), and SRR will differ from the SSIM index by an amount that depends on the difference between curves b and c of Fig. 3.

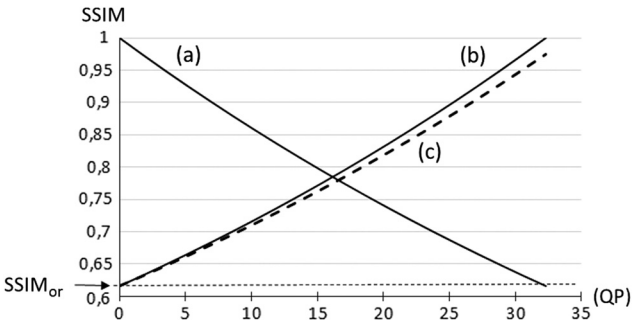

Fig. 3 Variation of: (a) $SSIM_{ot}$, (b) ideal $SSIM_{tr}$, and (c) real conditions $SSIM_{tr}$ versus QP.

4 Performance Evaluation

4.1 *Performance Comparison of Structural Similarity Reduced Reference to Structural Similarity*

The performance of the proposed method is evaluated by comparing SRR with the original SSIM index ($SSIM_{ot}$) for a large number of video frames. In this respect, a wide range of video sets were selected, which include 40 video sequences of various length, resolution, and content. The selected video sequences include 11 reference video sequences,[39] 2 long-duration sequences (Bigbuckbunny and Elephantsdream), and 27 nonreference video sequences retrieved from movie trailers. The total number of unique frames which were used for evaluating the proposed method is 60,866.

The original uncompressed video sequences were encoded at three QP values: 12, 22, and 32, which satisfactorily cover the achieved video quality range of the encoded/compressed video signals. Higher QP values were not examined because they lead to unacceptable video quality.[36] An initial qualitative comparison between SRR and SSIM index is shown in Fig. 4, which shows the variation of SRR and SSIM index for each frame of a video sequence (Kristen&Sara) with QP = 32. From Fig. 4, the qualitative similarity of RSS and SSIM index is obvious.

For the quantitative measurement of the performance of the proposed method, the MAPD for each frame i between SRR and SSIM index is calculated. MAPD is a widely used metric for measurement of the accuracy of a prediction method, specifically in trend estimation such as the the proposed method

$$MAPD = \frac{1}{n}\sum_{i=1}^{n}\frac{|SSIM_i - PredictedSSIM_i|}{SSIM_i}, \tag{12}$$

where $SSIM_i$ is the SSIM index per frame i and $PredictedSSIM_t$ is the SRR value for the frame t, according to Eq. (7).

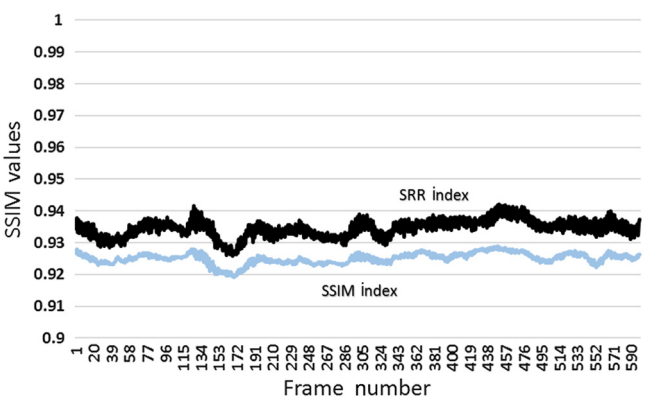

Fig. 4 Qualitative comparison of RSS and SSIM index applied on Kristen&Sara video sequence with QP = 32.

Table 2 MAPD for SRR and SSIM index for a white video reference pattern.

Test signal	Resolution	Frames	QP = 12	QP = 22	QP = 32
Apocalypto1	352 × 288	990	0.003053	0.005581	0.033561
Apocalypto2	352 × 288	990	0.004793	0.005936	0.020171
Apocalypto3	352 × 288	990	0.005903	0.007334	0.021447
Apocalypto4	352 × 288	501	0.005136	0.009094	0.017331
Mission1	352 × 288	990	0.005578	0.007233	0.016888
Mission2	352 × 288	990	0.006228	0.006695	0.019181
Mission3	352 × 288	293	0.004423	0.005531	0.016902
Superman1	352 × 288	990	0.004448	0.006062	0.019283
Superman2	352 × 288	990	0.00459	0.0069	0.021657
Superman3	352 × 288	268	0.001328	0.003436	0.035995
Insideman1	352 × 288	990	0.003798	0.004191	0.019316
Insideman2	352 × 288	990	0.006988	0.008945	0.012651
Insideman3	352 × 288	990	0.00526	0.006288	0.012146
Insideman4	352 × 288	376	0.00125	0.001084	0.027495
Davinci1	352 × 288	990	0.0029	0.003661	0.014045
Davinci2	352 × 288	990	0.005846	0.007804	0.019533
Davinci3	352 × 288	990	0.005416	0.005909	0.012271
Davinci4	352 × 288	627	0.007838	0.009838	0.018827
Basic1	352 × 288	990	0.005603	0.006751	0.015259
Basic2	352 × 288	990	0.00652	0.009051	0.015006
Basic3	352 × 288	990	0.006734	0.00765	0.014208
Basic4	352 × 288	351	0.003842	0.004624	0.010905
16blocks1	352 × 288	990	0.004517	0.004992	0.017711
16block2	352 × 288	990	0.006962	0.006389	0.012574
16block3	352 × 288	990	0.005295	0.005044	0.01225
16block4	352 × 288	451	0.003401	0.00297	0.007796
Batman1	352 × 288	2659	0.010258	0.014053	0.042584
Batman2	352 × 288	913	0.00619	0.011052	0.070104
Bigbuckbunny	640 × 360	14315	0.01091	0.020527	0.039061
Elephantsdream	640 × 360	15691	0.008501	0.013748	0.040883
Basketballpass	416 × 240	501	0.011807	0.007371	0.022277

Table 2 (*Continued*).

Test signal	Resolution	Frames	QP = 12	QP = 22	QP = 32
Bqsquare	416 × 240	601	0.006849	0.028219	0.101372
Bubbles	416 × 240	501	0.005385	0.015653	0.069205
Basketballdrill	832 × 480	501	0.016819	0.00973	0.0052
Bqmall	832 × 480	601	0.008256	0.005319	0.009498
Racehorses	832 × 480	300	0.009953	0.00794	0.022274
Partyscene	832 × 480	501	0.005572	0.01613	0.075519
Stockholm	1280 × 720	604	0.000514	0.030779	0.047909
Kristen&Sara	1280 × 720	600	0.008666	0.006363	0.010176
Foupeople	1280 × 720	600	0.009949	0.001049	0.001995
Mean value			0.00618198	0.00867315	0.02556165

Table 2 presents the MAPD for the experimental set of the 40 video sequences at the three QP values (i.e., 12, 22, and 32) and the mean value of MAPD for each QP.

Table 2 shows that the accuracy of the proposed method ranges from 0.62% (QP = 12) to 2.56% (QP = 32), which represents the worst case performance, showing that the proposed SRR method maintains a satisfactory performance across all the potential range of QP values, although better accuracy is achieved at lower QP values. The comparison between the correlation of QP and ground truth (i.e., MOS) versus the correlation of SRR scores and ground truth provides the result of 28.65, showing the advantage and the better performance of the proposed metric.

The experimental variation of MAPD versus QP is shown in Fig. 5(a), where the trend line is the dashed line. For comparison reasons, the theoretical MAPD is also shown in Fig. 5(b).

It is calculated from Eq. (12), where SSIM_i equals $\mathrm{SSIM}_{\mathrm{ot}}(\mathrm{QP})$, as calculated from Eq. (8). Also, $\mathrm{PredictedSSIM}_t$ equals SRR, i.e.,

$$\mathrm{SRR} = \frac{\mathrm{SSIM}_{\mathrm{or}}(\mathrm{QP})}{\mathrm{SSIM}_{\mathrm{tr}}(\mathrm{QP})}, \tag{13}$$

where $\mathrm{SSIM}_{\mathrm{tr}}(\mathrm{QP})$ corresponds to curve c of Fig. 3, which refers to an exemplary practical case.

The slopes of the two lines of Figs. 5(a) and 5(b) are the same, which shows that the theoretically calculated SRR is very close to the experimental results.

4.2 *Performance Comparison of Structural Similarity Reduced Reference to Subjective Difference Mean Opinion Scores and Other Assessment Methods*

According to the video quality experts group research,[40] in order to obtain a linear relationship between an objective assessment method score and its corresponding subjective score, each metric score x is mapped to $q(x)$. The nonlinear best-fitting logistic function $q(x)$ is given as follows:

$$q(x) = \beta_1 \left\{ \frac{1}{2} - \frac{1}{1 + \exp[\beta_2 (x - \beta_3)]} \right\} + \beta_4 \chi + \beta_5. \tag{14}$$

The parameters (β_1, β_2, β_3, β_4, and β_5) are calculated through minimizing the sum of squared differences among the subjective and the mapped scores. To compare the performance of a newly proposed SRR method with the existing

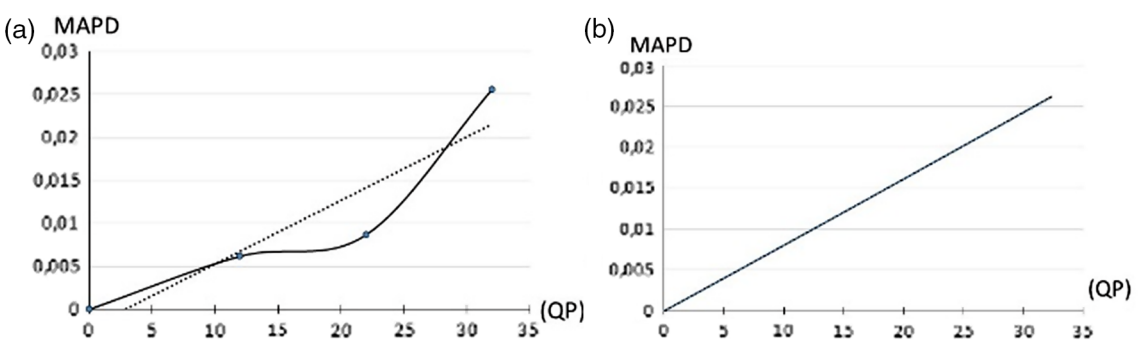

Fig. 5 Variation of MAPD versus QP: (a) experimental and (b) theoretical.

ones, performance evaluation metrics are used, such as the Pearson's linear correlation coefficient (LCC), which is the LCC between the predicted MOS and subjective MOS. LCC is a measure of prediction accuracy of an objective assessment metric, i.e., the capability of the metric to predict the subjective scores with low error. The LCC can be calculated using the below equation

$$\mathrm{LCC} = \frac{\sum_{i=1}^{M_d}(q_i - \bar{q})(s_i - \bar{s})}{[\sum_{i=1}^{M_d}(q_i - \bar{q})^2]^{\frac{1}{2}}[\sum_{i=1}^{M_d}(s_i - \bar{s})^2]^{\frac{1}{2}}}, \quad (15)$$

where s_i and q_i are the subjective and the mapped scores for the i'th frame of a video of size M_d, respectively, and $\bar{s}$ and $\bar{q}$ are the means of the mapped and subjective scores, respectively. A good objective assessment metric is expected to have high LCC (close to 1) in contrast to MAPD, which should have low values (i.e., close to 0), as shown in Sec. 4.1.

Moreover, the Spearman rank order correlation coefficient (SROCC), which measures the monotonicity of the proposed method against subjective human scores, was also applied. SROCC is a nonparametric measure of statistical dependence between two variables, which assesses how well the relationship between two variables can be described using a monotonic function. If there are no repeated data values, a perfect Spearman correlation (equal to 1) occurs when each of the variables is a perfect monotone function of the other.

Therefore, in order to evaluate the performance of the SRR method and derive the LCC, following the aforementioned methodology, the Laboratory for Image and Video Engineering (LIVE) video quality dataset[6,41] was used. The LIVE video quality database (VQDB) uses 10 uncompressed high-quality videos with a wide variety of content as reference videos. A set of 150 distorted videos were created from these reference videos (15 distorted videos per reference) using H.264-based compression. Then each degraded video in the LIVE VQDB was assessed by 38 human subjects in a single stimulus study with hidden reference removal, where the subjects scored the video quality on a continuous quality scale. The mean and variance of the difference mean opinion scores (DMOS) were obtained from the subjective evaluations.

A scatter plot of proposed objective SRR scores versus DMOS for all H.264 videos in the LIVE VQDB is shown in Fig. 6 along with the best-fitting logistic function.

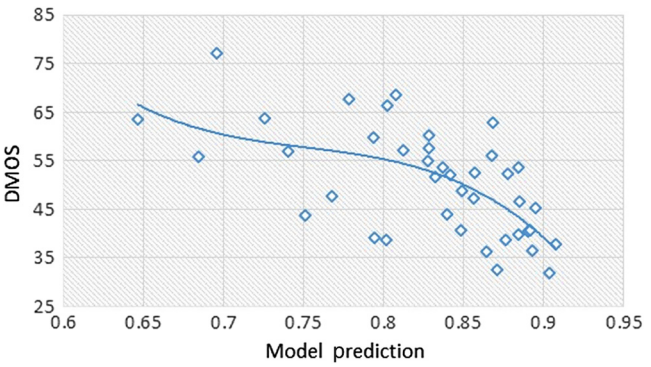

Fig. 6 Scatter plot of objective RSS scores versus DMOS with the best-fitting logistic function.

Table 3 Comparison of the performance of VQA algorithms—LCC and SROCC.

VQA method	Type	LCC	SROCC
RR-LHS[42]	RR	0.4557	0.4082
J.246[33]	RR	0.4488	0.4157
Peak signal-to-noise ratio (PSNR)	FR	0.5493	0.4585
Yang's RR VQA[43]	RR	0.5654	0.5366
Visual signal-to-noise ratio (VSNR)[44]	FR	0.6216	0.6460
Proposed SRR method	RR	0.6260	0.5862
Video quality metric (VQM)[34]	FR	0.6459	0.6520
SSIM[4]	FR	0.6656	0.6514
RR metric[35]	RR	0.7567	0.7486

The SROCC and the LCC are computed between the objective and the subjective scores. Table 3 shows the performance of the proposed SRR model against other VQA methods, both FR and RR, in terms of the SROCC and LCC.

According to Table 3, the accuracy of the proposed SRR method is measured more accurately than the RR VQA methods RR-LHS,[42] J.246,[33] and Yang's RR VQA [43] both in terms of LCC and monotonicity (SROCC), except for the RR metric,[35] which provides better results. Similarly, the accuracy of the proposed SRR method is better than those of the PSNR and VSNR FR VQA methods in terms of LCC, while it is slightly lower than the performance of VQM and SSIM index, as expected, due to the reduced reference nature of the proposed methodology. In terms of monotonicity (SROCC), the proposed method performs better than the PSNR, but lower than the rest of the VQA methods, without, however, significantly deviating from their performance range.

5 Conclusions

This paper proposes an RR VQA method using SSIM index as a tool to extract features from both the original and the target video sequences, using a reference video pattern. The method is suitable for monitoring the video quality in real time and across the service provision chain. The performance of the proposed method was evaluated using a large experimental set of 40 reference and nonreference video sequences, with spatial resolution ranging from common intermediate format up to high definition, utilizing a static white video as a relative reference pattern. The proposed method maintains a satisfactory performance across all the potential range of QP values, although better accuracy is achieved at lower QP values. Moreover, comparison to subjectively evaluated scores of LIVE video quality dataset shows that the accuracy of the proposed method is better than the average performance of RR VQA methods and is within the performance range of the FR VQA methods. Finally, another advantage of the proposed method is the low bit rate reference information signal that must

be sent to the end user, which ranges between 400 and 600 bps.

Acknowledgments

This work was undertaken under the Information Communication Technologies (FP7-ICT-2013-11) EU FP7 T-NOVA project, which is funded by the European Commission under the Grant No. 619520.

References

1. H. Koumaras, M. A. Kourtis, and D. Martakos, "Benchmarking the encoding efficiency of H.265/HEVC and H.264/AVC," in *Future Network & Mobile Summit 2012*, 4–6 July 2012, Berlin, Germany (2012).
2. H. Koumaras et al., "Impact of H.264 advanced video coding interframe block sizes on video quality," in *VISAPP 2012*, 24–26 February 2012, Rome, Italy (2012).
3. F. Liberal et al., "Multimedia content delivery in SDN and NFV-based towards 5G networks," *IEEE COMSOC MMTC E-Lett.* **10**(4), 6–10 (2015).
4. Z. Wang et al., "Image quality assessment: from error visibility to structural similarity," *IEEE Trans. Image Process.* **13**(4), 600–612 (2004).
5. H. Koumaras et al., "A framework for end-to-end video quality prediction of MPEG video," *J. Visual Commun. Image Represent.* **21**, 139–154 (2010).
6. K. Seshadrinathan et al., "Study of subjective and objective quality assessment of video," *IEEE Trans. Image Process.* **19**(6), 1427–1441 (2010).
7. Z. Wang and E. P. Simoncelli, "Reduced-reference image quality assessment using a wavelet-domain natural image statistic model," *Proc. SPIE* **5666**, 149 (2005).
8. Q. Li and Z. Wang, "Reduced-reference image quality assessment using divisive normalization-based image representation," *IEEE J. Sel. Top. Signal Process.* **3**(2), 201–211 (2009).
9. T. M. Cover and J. A. Thomas, *Element of Information Theory*, Wiley, New York, (1991).
10. L. Ma et al., "Reduced-reference image quality assessment using reorganized DCT-based image representation," *IEEE Trans. Multimedia* **13**(4), 824–829 (2011).
11. R. Soundararajan and A. C. Bovik, "RRED indices: reduced reference entropic differencing framework for image quality assessment," in *Proc. IEEE Int. Conf. Acoustics, Speech and Signal Processing*, pp. 1149–1152 (2011).
12. J. A. Redi et al., "Color distribution information for reduced-reference assessment of perceived image quality," *IEEE Trans. Circuits Syst. Video Technol.* **20**(12), 1757–1769 (2010).
13. K. Zeng and Z. Wang, "Temporal motion smoothness measurement for reduced-reference video quality assessment," in *Proc. IEEE Int. Conf. Acoustics, Speech and Signal Processing*, pp. 1010–1013 (2010).
14. K. Zeng and Z. Wang, "Quality-aware video based on robust embedding of intra and interframe reduced-reference features," in *Proc. IEEE Int. Conf. Image Processing*, pp. 3229–3232 (2010).
15. P. Le Callet, C. V. Gaudin, and D. Barba, "Continuous quality assessment of MPEG2 video with reduced reference," in *Proc. Int. Workshop on Video Processing Quality Metrics for Consumer Electronics*, pp. 11–16 (2005).
16. Z. Wang, H. R. Sheihk, and A. C. Bovik, "Objective video quality assessment," in *The Handbook of Video Databases: Design and Application*, B. Furht and O. Marqure, Eds., pp. 1041–1078, CRC Press, Boca Raton, Florida (2003).
17. Z. Wang, L. Lu, and A. C. Bovik, "Video quality assessment based on structural distortion measurement," *Signal Process. Image Commun.* **19**(2), 121–132 (2004).
18. I. P. Gunawan and M. Ghanbari, "Reduced-reference picture quality estimation by using local harmonic amplitude information," in *Proc. London Communications Symp. 2003*, pp. 137–140, University College London, United Kingdom (2003).
19. I. Gunawan and M. Ghanbari, "Reduced-reference video quality assessment using discriminative local harmonic strength with motion consideration," *IEEE Trans. Circuits Syst. Video Technol.* **18**, 71–83 (2008).
20. Y. K. Wang et al., "Supervised automatic detection of UWB ground-penetrating radar targets using the regression SSIM measure," *IEEE Geosci. Remote Sens. Lett.* **13**(5), 621–625 (2016).
21. T.-C. Chang, S. S.-D. Xu, and S.-F. Su, "SSIM-based quality-on-demand energy-saving schemes for OLED displays," *IEEE Trans. Syst. Man Cybern. Syst.* **46**(5), 623–635 (2016).
22. Y. Fang et al., "Optimized multioperator image retargeting based on perceptual similarity measure," *IEEE Trans. Syst. Man Cybern. Syst.* **99**, 1–11 (2016).
23. A. Webster et al., "An objective video quality assessment system based on human perception," *Proc. SPIE* **1913**, 15 (1993).
24. S. Wolf and M. H. Pinson, "Spatial-temporal distortion metric for in-service quality monitoring of any digital video system," *Proc. SPIE* **3845**, 266–277 (1999).
25. S. Wolf and M. H. Pinson, "Low bandwidth reduced reference video quality monitoring system," in *Proc. Int. Workshop on Video Processing Quality Metrics for Consumer Electronics*, pp. 76–79 (2005).
26. T. Kusuma and H.-J. Zepernick, "A reduced-reference perceptual quality metric for in-service image quality assessment," in *Joint First Workshop on Mobile Future and Symp. on Trends in Communications (SympoTIC '03)*, pp. 71–74 (2003).
27. I. P. Gunawan and M. Ghanbari, "Reduced reference picture quality estimation by using local harmonic amplitude information," in *Proc. London Communications Symp.*, pp. 137–140 (2003).
28. K. Chono et al., "Reduced-reference image quality assessment using distributed source coding," in *2008 IEEE Int. Conf. on Multimedia and Expo*, pp. 609–612 (2008).
29. M. Carnec, P. Le Callet, and D. Barba, "An image quality assessment method based on perception of structural information," in *Proc. 2003 Int. Conf. on Image Processing (ICIP '03)*, Vol. 3, pp. 185–188 (2003).
30. M. Carnec, P. Le Callet, and D. Barba, "Visual features for image quality assessment with reduced reference," in *Proc. IEEE Int. Conf. Image Processing*, Vol. 1, pp. 421–424 (2005).
31. H. B. Barlow, "Possible principles underlying the transformation of sensory messages," in *Sensory Communication*, W. A. Rosenblith, Ed., pp. 217–234, MIT Press, Cambridge, Massachusetts (1961).
32. E. P. Simoncelli and B. Olshausen, "Natural image statistics and neural representation," *Annu. Rev. Neurosci.* **24**, 1193–1216 (2001).
33. ITU-T, "Perceptual visual quality measurement techniques for multimedia services over digital cable television networks in the presence of a reduced bandwidth reference," ITU-T Rec. J.246, 2008, http://www.itu.int/rec/T-REC-J.246/en (22 June 2016).
34. M. H. Pinson and S. Wolf, "A new standardized method for objectively measuring video quality," *IEEE Trans. Broadcast.* **50**(3), 312–322 (2004).
35. L. Ma, S. Li, and K. N. Ngan, "Reduced-reference video quality assessment of compressed video sequences," *IEEE Trans. Circuits Syst. Video Technol.* **22**(10), 1441–1456 (2012).
36. M. Razaak, M. Martini, and K. Savino, "A study on quality assessment for medical ultrasound video compressed via HEVC," *IEEE J. Biomed. Health Inform.* **18**(5), 1552–1559 (2014).
37. H. Koumaras, "Method for predicting perceived quality of encoded video in relation to encoding bit rate and spatiotemporal content dynamics," PhD Dissertation, University of Athens, Computer Science and Telecommunications Department, Athens, Greece (2007).
38. H. Koumaras et al., "Quantified PQoS assessment based on fast estimation of the spatial and temporal activity level," *Multimedia Tools Appl.* **34**(3), 355–374 (2007).
39. Hannover University, ftp://ftp.tnt.uni-hannover.de (22 June 2016).
40. Video quality expert group (VQEG), "Final report from the video quality experts group on the validation of objective models of video quality assessment II," http://www.vqeg.org (22 June 2016).
41. K. Seshadrinathan et al., "A subjective study to evaluate video quality assessment algorithms," *Proc. SPIE* **7527**, 75270H (2010).
42. I. P. Gunawan and M. Ghanbari, "Reduced-reference video quality assessment using discriminative local harmonic strength with motion consideration," *IEEE Trans. Circuits Syst. Video Technol.* **18**(1), 71–83 (2008).
43. S. Yang, "Reduced reference MPEG-2 picture quality measure based on ratio of DCT coefficients," *Electron. Lett.* **47**(6), 382–383 (2011).
44. D. M. Chandler and S. S. Hemami, "VSNR: a wavelet-based visual signal-to-noise ratio for natural images," *IEEE Trans. Image Process.* **16**(9), 2284–2298 (2007).

Michail-Alexandros Kourtis received his diploma and master's degree in computer science from Athens University of Economics and Business in 2011 and 2014, respectively. Since January 2010, he has been working at the National Center of Scientific Research (NCSR) "Demokritos" in various research projects. Currently, he is pursuing his PhD at the University of the Basque Country [universidad de pais vasco/Euskal Herriko Unibertsitatea (UPV/EHU)]. His research interests include video processing, video quality assessment, image processing, network function virtualization, and software defined.

Harilaos Koumaras received his BSc degree in physics in 2002, his MSc degree in electronic automation and information systems in 2004 and his PhD in 2007 in computer science, all awarded from the University of Athens. He is a research associate at the NCSR "Demokritos," with publications at international conferences, journals, and book chapters. Currently, he is the author or coauthor of more than 80 scientific papers, numbering at least 350 nonself-citations.

Fidel Liberal received his MSc degree in telecommunication engineering in 2001. He received his PhD from the University of the Basque Country (UPV/EHU) in 2005 for his work in the area of holistic management of quality in telecommunications services. He has coauthored more than 80 conference and journal papers with more than 300 citations. His current research interests include quality management and multicriteria optimization for multimedia services in 5G.