Received 7 September 2015
Revised 5 October 2015
Accepted 1 February 2016



# An experimental survey of no-reference video quality assessment methods

Maria Torres Vega
*Department of Electrical Engineering, Einhoven University of Technology, Eindhoven, The Netherlands*
Vittorio Sguazzo
*Universita degli Studi di Salerno, Fisciano, Italy*
Decebal Constantin Mocanu
*Department of Electrical Engineering, Technische Universiteit Eindhoven, Eindhoven, The Netherlands, and*
Antonio Liotta
*Department of Electrical Engineering, Eindhoven University of Technology, Eindhoven, The Netherlands*

## Abstract

**Purpose** – The Video Quality Metric (VQM) is one of the most used objective methods to assess video quality, because of its high correlation with the human visual system (HVS). VQM is, however, not viable in real-time deployments such as mobile streaming, not only due to its high computational demands but also because, as a Full Reference (FR) metric, it requires both the original video and its impaired counterpart. In contrast, No Reference (NR) objective algorithms operate directly on the impaired video and are considerably faster but loose out in accuracy. The purpose of this paper is to study how differently NR metrics perform in the presence of network impairments.

**Design/methodology/approach** – The authors assess eight NR metrics, alongside a lightweight FR metric, using VQM as benchmark in a self-developed network-impaired video data set. This paper covers a range of methods, a diverse set of video types and encoding conditions and a variety of network impairment test-cases.

**Findings** – The authors show the extent by which packet loss affects different video types, correlating the accuracy of NR metrics to the FR benchmark. This paper helps identifying the conditions under which simple metrics may be used effectively and indicates an avenue to control the quality of streaming systems.

**Originality/value** – Most studies in literature have focused on assessing streams that are either unaffected by the network (e.g. looking at the effects of video compression algorithms) or are affected by synthetic network impairments (i.e. via simulated network conditions). The authors show that when streams are affected by real network conditions, assessing Quality of Experience becomes even harder, as the existing metrics perform poorly.

**Keywords** Network impaired videos, No-reference video quality, Quality of experience

**Paper type** Research paper

International Journal of Pervasive Computing and Communications
Vol. 12 No. 1, 2016
pp. 66-86

1742-7371
DOI 10.1108/IJPCC-01-2016-0008

## 1. Introduction

Video streams are affected by network protocols and impairments (such as jitter and packet loss) according to non-linear quality degradation functions, often in unexpected

and unpredictable ways (Liotta, 2013). Thus, assessing video degradation has become the subject of many studies, especially by means of Quality of Experience (QoE) tools (Menkovski *et al.*, 2010). What is really crucial is to find reliable yet simple and scalable methods that can be used by service and content providers to manage their services, adjusting the streams' quality according to both the network conditions and the human perceptual models. This goal becomes particularly challenging, as quality assessment algorithms are either too complex or too unreliable. Herein, we experimentally survey the most popular algorithms, identifying their operational conditions.

QoE is defined as the degree of delight or annoyance of the user of an application or service (Le Callet *et al.*, 2012). Due to its subjective essence, the legitimate judges of visual quality are the humans, whose opinion can be obtained through subjective analyses (Shahid *et al.*, 2014). In practice, presented stimuli (e.g. impaired video sequences) are rated by subjects under controlled conditions (Zinner *et al.*, 2010). These ratings express the subjective QoE (sQoE) described typically by the Mean Opinion Score (MOS). However, due to the time-consuming nature and bias of subjective experiments (Menkovski *et al.*, 2011), in the past years, great effort has been placed onto developing objective quality metrics which could provide with a valid alternative, that is, objective QoE (oQoE) (Staehle *et al.*, 2008).

The ultimate goal of the oQoE metrics is to provide the best possible correlation to subjective studies and the human vision system (HVS) by means of only the reference (original) and the received material. Depending on the amount of reference information necessary to perform the assessment, the oQoE approaches are classified in three categories: Full Reference (FR), Reduced Reference (RR) and No Reference (NR). FR and RR metrics require the original material (either in its totality or through the analysis of certain features) to perform their assessment. Examples of these metrics are the FR Peak-Signal to Noise Ratio (PSNR) and the Structural Similarities (SSIM) algorithms (Wang *et al.*, 2004) or the RR approaches of Mocanu *et al.* (2014b, 2015). Specifically because of its good correlation with subjective values, the Video Quality Metric (VQM) (Pinson and Wolf, 2004) is commonly used as the FR benchmark metric (Chikkerur *et al.*, 2011). However, its high complexity, running time and the fact that, as a FR metric, it requires both the original and the impaired material make VQM not a viable solution in real-time deployments, such as mobile streaming services. To fill the gap, NR metrics have started to take a predominant role. Their biggest asset is the fact that they do not rely on highly complex comparisons among streams but on the measurements of external factors to model the multimedia quality. These metrics, which base their quality assessment on the analysis of video features at the bit-stream level (bitrate, scene complexity, video motion and other parameters in the codec) or at the frame pixel level (blur, noise, blockiness), provide a very fast or even real-time assessment while being able to be deployed in lightweight environments (Torres Vega *et al.*, 2015a, 2015b). Despite this fact, their accuracy in assessing video degradations and their correlation to subjective analysis are still open issues. This situation makes it particularly hard to automate the assessment of real-time streams that have been subjected to network packet loss, which is the main focus of our study. Most studies in the literature have focused on the assessment of streams that are either unaffected by the network (e.g. looking at the effects of video compression algorithms) or are affected by synthetic network impairments (i.e. via simulated network conditions). We show that when

streams are affected by real network conditions, assessing QoE becomes even harder, as the existing metrics perform poorly.

In this paper, we present an experimental analysis of the accuracy of NR metrics in network impaired videos, extending the early results in our previous work (Torres Vega *et al.*, 2015c). Based on a prior thorough study on the currently used state-of-the-art NR metrics and the features most commonly used for their assessment, we designed and implemented an RTP-video client tool which first analyses the impaired video using eight different NR features alongside the lightweight FR metric SSIM and VQM (used as the quality benchmark). In an experimental test bed, we prepared an impaired video set of 960 videos, which covers a diverse set of video types, encoding conditions and network situations, and we deployed our methodology and explored the accuracy of the NR features and SSIM by correlating their values with the benchmark quality. The purpose of this study is to provide a NR framework for video quality assessment and to show the extent by which network impairments affect different video types, correlating the accuracy of NR metrics to the FR benchmark. Our work helps identify the conditions under which simple metrics may be used effectively and in line with human perception.

The remainder of this paper is organized as follows. Section 2 provides an overview on the video features currently used for assessing the quality in a NR manner. Section 3 presents our proposed methodology, as well as giving insights on the development of the different NR metrics. The experimental video set is introduced in Section 4. Section 5 provides an evaluation of results focusing not only on impairments derived from the compression degradation but also the ones derived from real network conditions. Finally, Section 6 draws conclusions, highlighting key contributions and suggesting directions for future work.

## 2. NR-Video features and artifacts

Reibman *et al.* (2005) classified NR approaches as either stemming from statistics derived from pixel-based features, NR pixel (NR-P), or computed directly from the coded bitstream, NR bitstream (NR-B). In a more recent classification, Shahid *et al.* (2014) added to this classification a third category in which approaches combining both pixel and bitstream assessments are included, that is, Hybrid NR-P-B metrics.

NR-P methods have focused their attention on the employment of certain artifacts related to a specific type of degradation of the visual quality. Blur, noise, blockiness or temporal impairments have been quantified for measuring the end-user's quality. Blur, measured frame by frame, appears as a loss of spatial detail and a reduction of edge sharpness (Winkler, 2005). Examples of blur-based NR video quality assessment can be found in Ciancio *et al.* (2011) and Ferzli and Karam (2006). Noise has also been used to assess quality, like in the block-based approach of Rank *et al.* (1999). Blockiness manifests itself as a discontinuity between adjacent blocks in images and video frames (Hemami and Reibman, 2010). Several research lines have focused on blockiness, examples of that are the Block-Edge metric of Wu and Yuen (1997) or the HVS-based blocking method of Liu and Heynderickx (2008). Finally, temporal impairments incur by the network through delay or packet loss. These result in a degradation of the video in the form of jerkiness (non-fluent and non-smooth presentation of frames) (Borer, 2010), frame freezes or jitter. Because videos can be affected by more than one artifact at a time, methods combining contributions of different artifacts appeared, such as the linear combination of noise and blur components of Choi *et al.* (2009).

NR-B methods are relatively simpler to compute, and quality scores can be obtained in the absence of the full decoder. But they tend to have a limited scope of application, as they are usually designed for specific coding techniques, for example, H.264/AVC (Brandão and Queluz, 2010). Another issue is to find the correlation between the bitlayer parameters and quality, without increasing the complexity. Learning tools have proven to be a promising solution for this type of approaches. Shahid *et al.* (2011) proposed a model combining different bitstream-layer features using an Artificial Neural Network to estimate the quality.

The performance of NR-B metrics can be enhanced by adding some input form of NR-P-based quality assessment. These methods are called hybrid methods. Shanableh (2011) suggest a multi-pass prediction system based on step-wise regression using features included in the coding information of a Macro Block (MB), some relative measures of motion vector of neighboring MBs and some numerical values related to textures of the MB. This method has shown good correlation to SSIM, but its complexity makes it not quite fit for online analysis. Another interesting approach is the one provided by Keimel *et al.* (2012). They measure the quality by linearly combining bitstream and pixel-related features. Despite substantially reducing complexity, their approach does not, however, correlate well with either PSNR (Peak Signal to Noise Ratio) or SSIM, two of the state-of-the-art objective metrics in current use. Finally, in our previous research, we developed a lightweight algorithm combining bitstream parameters (video bitrate, complexity and motion) with pixel artifacts (blur and noise). In Torres Vega *et al.* (2015a), we presented the algorithm and showed a high correlation with SSIM.

## 3. Methodology

The aim of this study is to understand how accurately different NR features and artifacts assess degradations on the video quality. Thus, the first step was to define which metrics would be used. Then, we had to decide which FR metric would be most suited as benchmarking and how to perform it.

As already introduced in the previous section, NR metrics have been traditionally classified according to the features they take into consideration to assess multimedia quality. In this way, NR-P metrics focus on frame- and pixel-level features and NR-B on the bitstream characteristics. Both types combined have been demonstrated to provide better analysis and are, nowadays, the base for the development of the state-of-the-art in NR-metrics. Following this trend, we selected metrics in both levels. Because the purpose of this analysis is to understand under which circumstances a very high computational FR analysis could be substituted by a low complexity NR metric, an extra requirement for all the NR features selected was for them to be obtainable in real time, at low computational complexity and, thus, suited both for standard and very lightweight devices.

A video stream can be characterized by several parameters that will affect video types differently. Parameters regarding the video composition have been demonstrated to affect quality to a large extent because they robustly combine different characteristics obtained from the video encoding. In particular, the scene complexity and the video motion have proven to provide a high level of correlation with quality degradations (Liotta *et al.*, 2013; Hu and Wildfeuer, 2009). First, the scene complexity quantifies the number of objects or elements present at the video scene. Second, the video motion can

be described as the amount of movement present in the video. Both features can be empirically derived from the received encoded video in real time following equations (1) and (2) for scene complexity and video motion, respectively (Liotta *et al.*, 2013):

$$C = \frac{Bits_I}{2 * 10^6 * 0.91^{QP_I}} \tag{1}$$

$$M = \frac{Bits_P}{2 * 10^6 * 0.87^{QP_P}} \tag{2}$$

where $Bits_I$ and $Bits_P$ are bits of coded Intra (I) and Inter (P) frames, and $QP_I$ and $QP_P$ represent the average I-Frames and P-Frames quantization parameter.

Extensive research has been performed on NR-pixel-level features. Metrics such as the level of noise, the clipping or the frame edges' blur have been used to measure multimedia quality in a NR way. For this analysis, we focused on four features which have been demonstrated to provide a degradation assessment and which are fundamental when dealing with impaired video streams. The network and the encoding will provoke the appearance of different artifacts on the frames. First of all, the images will develop blocks in which the image is not clear and cannot be processed. This effect is called blockiness and can be easily calculated for each of the video frames. In our study, we followed the procedure described in the study by Perra (2014) and Wu and Yuen (1997). Furthermore, degradations can be observed in terms of the sharpness and the cleanness of the frame. A blur and noise measure gives the quantitative assessment on both effects (Choi *et al.*, 2009; Torres Vega *et al.*, 2015a, 2015b, 2015c). Choi *et al.* (2009) developed an algorithm to assess the quality by means of blur and noise components. From this algorithm, four measurements can be obtained: average video blur mean, average blur ratio, average noise mean and average noise ratio. These four features are linearly combined to obtain the quality value. In our previous work (Torres Vega *et al.*, 2015a), we extended their algorithm to videos by introducing weights adapted to the video type. Here, the accuracy of each of these four features is assessed independently. The blockiness, noise (mean and ratio) and blur (mean and ratio) values are obtained per frame and averaged by the number of frames in the video. Finally, temporal aspects such as freezes and jerks can degrade the quality in unexpected and unrecoverable ways. In our set of features, we include temporal aspects by measuring the video jerkiness, which can be derived by means of the minimum square difference (MSD) between frames method developed by Borer (2010). Thus, this calculation is performed in sequential pairs of frames all over the video and averaged and normalized at video ending.

Objective QoE FR metrics have been traditionally used as a valid alternative to assess end-user quality in the absence of subjective data. A well-known example of these metrics is SSIM which, thanks to a very exhaustive frame-by-frame analysis, provides results in line with the HVS. However, this algorithm, although well suited for images, lacks a temporal assessment, and thus, it fails to perform accurately for the assessment of videos. VQM (Pinson and Wolf, 2004), on the other hand, combines analyses both in the temporal and the spatial video levels to obtain an assessment highly correlated with the HVS (Chikkerur *et al.*, 2011). Thus, we used VQM as the baseline benchmark for the NR metrics. Furthermore, to fully understand how far in terms of accuracy is SSIM

compared to the state-of-the-art, we decided to include this lightweight FR metric in the accuracy framework.

Finally, to automate the whole measurement, calculation and benchmark, we designed an RTP-Video Client tool as schematized in Figure 1. On arrival to the client, the eight NR-features present in the impaired video are quantified. At the same time, a copy of the original video is made available on the client side to be used to perform two FR assessments, SSIM and VQM. Once all metrics and the benchmark quality have been obtained, they are pre-processed to simplify the comparison. First, metrics are normalized between 0 and 1. Second, metrics measuring the level of degradation instead of the actual quality are inverted. As a consequence of this process, all metrics and the benchmark quality are in the same range between 0 and 1, where 0 represents total loss of quality and 1 means full quality, that is, no degradation. Accuracy is then assessed by means of a Pearson correlation (Kendall *et al.*, 1987) between the metric under scrutiny and the benchmark normalized quality.

## 4. Video data set

Figure 2 shows the test-bed used for generating our video quality evaluation framework. The RTP-video server streams videos on demand to the RTP-video client. Between server and client, a network emulator is installed. This device is able to emulate real-time network conditions. On reception of the impaired video, the client performs the NR-video

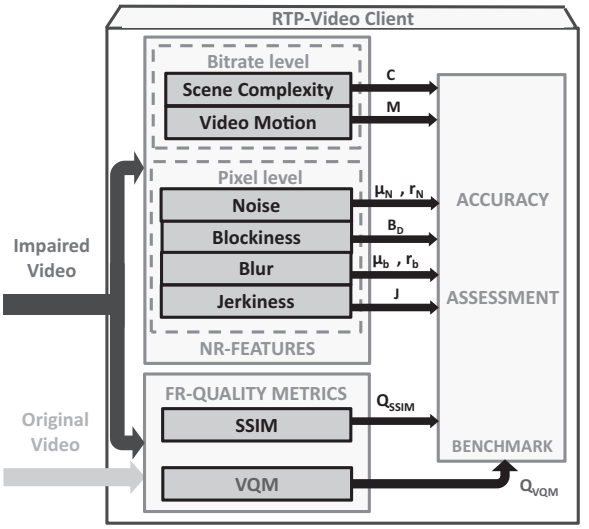


**Figure 1.**
NR metrics accuracy assessment tool

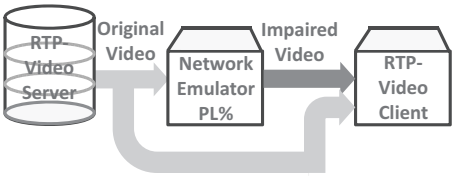


**Figure 2.**
Test-bed block diagram

features assessments and the FR metrics following the methodology introduced in the previous section.

The original video set used for the evaluation consists of 10-s videos at 25 fps from the Live Quality Video Database (Seshadrinathan *et al.*, 2010a, 2010b). Each of the videos is of a different dynamic composition and type (Table I). These videos are compressed to H264/MPEG4 at a resolution of 768 × 432 at eight different bitrates (64 kbps, 640 kbps, 768 kbps, 1,024 kbps, 2,048 kbps, 3,072 kbps, 4,096 kbps and 5,120 kbps). The selection of the encoding bitrates has been done in a way as to obtain the most diverse variety of video qualities. For example, very low quality transmissions (64 kbps) are nowadays, with the currently used systems and Internet speeds, highly unlikely to occur. However, they could still be used in mobile devices in a very congested network. With this variety of bitrates, our data set covers a broad range of video types, which allows a comprehensive NR analysis.

Packet loss has been demonstrated to be the main cause of degradation in RTP video transmissions (Mocanu *et al.*, 2014c; Suárez *et al.*, 2015). Thus, to make our data set most suitable to real network situations, we focused on videos impaired by the influence of packet losses. For the generation of the full data set, 80 original videos (ten types at eight quality levels each) were transmitted from server to client through a lossy network. Each video type and bitrate was streamed through the network 11 times. Each video was subjected to all different levels of packet loss (0.5, 1, 1.5, 2, 2.5, 3, 3.5, 4, 4.5, 5 and 10 per cent). This means that we generated a video set consisting of 960 videos, obtained from 10 original videos, each encoded at 8 different bitrates and 12 different conditions (1 compression degradation + 11 compression degradation and packet loss).

**Table I.**
Video test set: acronym, name and description

| Acronym | Name | Description |
|---|---|---|
| bs1 | Blue sky | Circular camera motion<br>Blue sky and trees |
| mc1 | Mobile calendar | Camera pan, horizontal tor train<br>Calendar moving vertically |
| pa1 | Pedestrian area | Still camera<br>People walking on intersection |
| pr1 | Park run | Camera pan<br>Person running across a park |
| rb1 | River bed | Still camera<br>River bed, pebbles in the water |
| rh1 | Rush hour | Still camera<br>Rush hour traffic on the street |
| sf1 | Sunflower | Still camera<br>Bee over a sunflower in close-up |
| sh1 | Shields | Camera pan, still and zoom<br>Person across a display of shields |
| st1 | Station | Still camera<br>Railway track, one train and people walking across |
| tr1 | Tractor | Camera pan<br>Tractor moving across fields |

## 5. Evaluation

For the sake of simplicity, we have divided the analysis in two different parts. In the first one (Section 5.1), we focus on the accuracy of NR metrics in videos that are affected only by compression degradation, testing the original 80 videos, compressed at different levels (no network impairments). This provides a clean benchmark for the different NR metrics under conventional conditions (no network effects). In the second section (Section 5.2), we extend the analysis to study how network impairments affect the accuracy of NR metrics, thus showing more realistic conditions than what is known from the literature (including the whole 960 video set).

### *5.1 Quality versus compression*

In this first part of the study, we focused on analyzing the accuracy of the NR-metrics when the videos are only affected by the degradations derived from the video compression. Thus, we used the original 80 videos compressed from the original 10 raw videos at 8 different bitrates. The results are shown in two figures. First, we show the behavior of the benchmark quality in these original 80 videos (Figure 3). And then, we present the accuracy results for the eight NR features and SSIM (Figure 4).

In the colormap presented in Figure 3, bitrates can be seen on the *x*-axis, while the main *y*-axis shows the 10 video types, and the secondary *y*-axis presents the average quality value per bitrate (aggregating all video types). Dark blue indicates full quality. As the quality degrades, the sample color goes from light blue to yellow, orange and

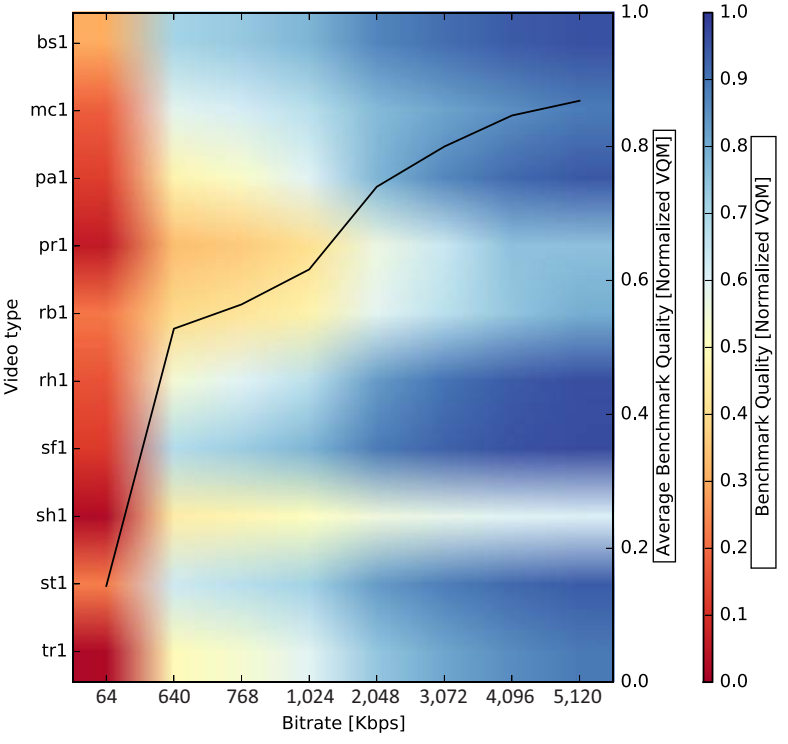


**Figure 3.**
Benchmark quality of the original 80 videos

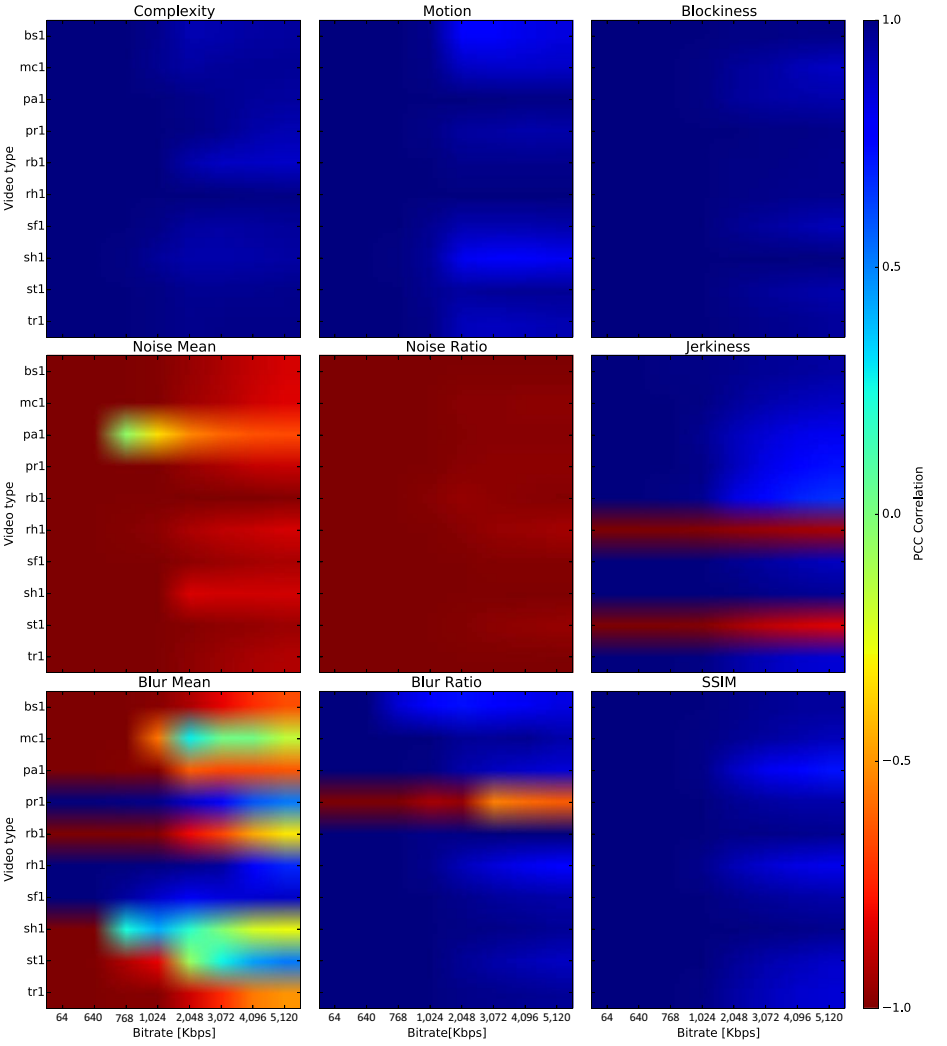


**Figure 4.**
NR metrics and SSIM correlation to the benchmark quality for all the videos for encoding-related degradation

finally red (poorest quality). As expected, for all video types, quality increases proportionally to the bitrate. Most of the video types show red or dark orange values at very low bitrates and dark blues at the highest ones. However, it can be seen that each video type behaves in a different manner. While some videos, such as the Rush hour (rh1, sixth line from the top) or the Sunflower (sf1, seventh line from the top), reach maximum quality at bitrates of 3 or 4 Mbps, others like the Park run (pr1, fourth line from the top), the River bed (rb1, fifth line from the top) or the Shields (sh1, eighth line) are still far from the maximum quality index (1, dark blue) encoded at 5 Mbps. These different behaviors demonstrate the generality of our data set.

The accuracy of the eight NR metrics and SSIM in the presence of compression degradation is shown in Figure 4. Each of the colormaps shows one metric for the 80

videos (type and bitrate). As for the previous figure, the *x*-axis presents the bitrate and the *y*-axis shows the ten video types. A dark blue value means full correlation (quality = 1). As the correlation degrades, the color goes from light blue to green (0 or no correlation), yellow and final red (−1, the two metrics are anti-correlated, that is, they follow opposite trends).

If we look at the bitstream layer parameters, complexity and motion (first and second colormaps, first row), then it can be seen that both features correlate very well for all the videos (dark blue). However, in some specific cases, such as the river bed (rb1, fifth line) – in the case of complexity – or the Park run (pr1, fourth line) – in the case of motion – the correlation starts degrading from a certain bitrate on. The reason for this is that at certain levels of compression (bitrate), the metric saturates and cannot further improve, while the benchmark quality (which takes into account many different parameters) continues improving.

The blockiness (third colormap, first row) shows full correlation (dark blue) for nearly all bitrates and video types. In some cases, the correlation degrades at very high bitrates. Examples of this behavior are the Pedestrian area (pa1, third row), the Sunflower (sf1, seventh row) or the Station (st1, ninth row). The reason for it can be found by looking at the reference quality of these videos (Figure 3). In these three cases, the quality is maximum from early stages of bitrate (roughly 2 Mbps), while the blockiness quality (Figure 4) keeps showing improvements.

While the noise components (first and second colormaps, second row) show no correlation with the benchmark quality for any of the videos or bitrates, the blur components (first and second colormaps, third row) of the videos show quite an interesting behavior. For seven of the videos, the average blur is highly anti-correlated, while the ratio is highly correlated. However, this cannot be considered a general characteristic. The video Park run (pr1, fourth row) presents the complete opposite pattern. Furthermore, the Rush hour (rh1) and the Sunflower (sf1), sixth and seventh lines, show full correlation in both metrics for all the bitrates.

The jerkiness shows high levels of correlation (dark blue) in eight of the ten video types. The anti-correlation seen for the Rush hour (rh1, sixth line) and the Station (st1, tenth line) can be explained by the video composition which makes them quite resistant to jerkiness artifacts.

Finally, SSIM (third colormap, third row) shows high correlation with the benchmark quality. It can be observed that in some of the cases, the correlation degrades lightly for higher bitrates, while in others, the correlation is maintained.

### *5.2 Quality versus network impairments*

In this second part of the study, we extended the analysis to the whole 960 video set. As in the previous case, the results are shown in two parts. First, we show the benchmark quality in the 960 videos of the data set (Figure 5). Then, we present the accuracy results for the eight NR features and SSIM (Figures 6-15) for all ten videos types.

Figure 5 presents the quality analysis of the impaired video set. Each of the colormaps shows a video type. The packet loss level is seen on the *x*-axis, while the main *y*-axis shows the bitrates and the secondary *y*-axis presents the average quality (aggregated across all bitrates) for each of the packet loss levels. As in the previous case, dark blue indicates full quality, and as the quality degrades, the sample color goes from light blue to yellow, orange and a final red (poorest quality). In all these cases, dark blue

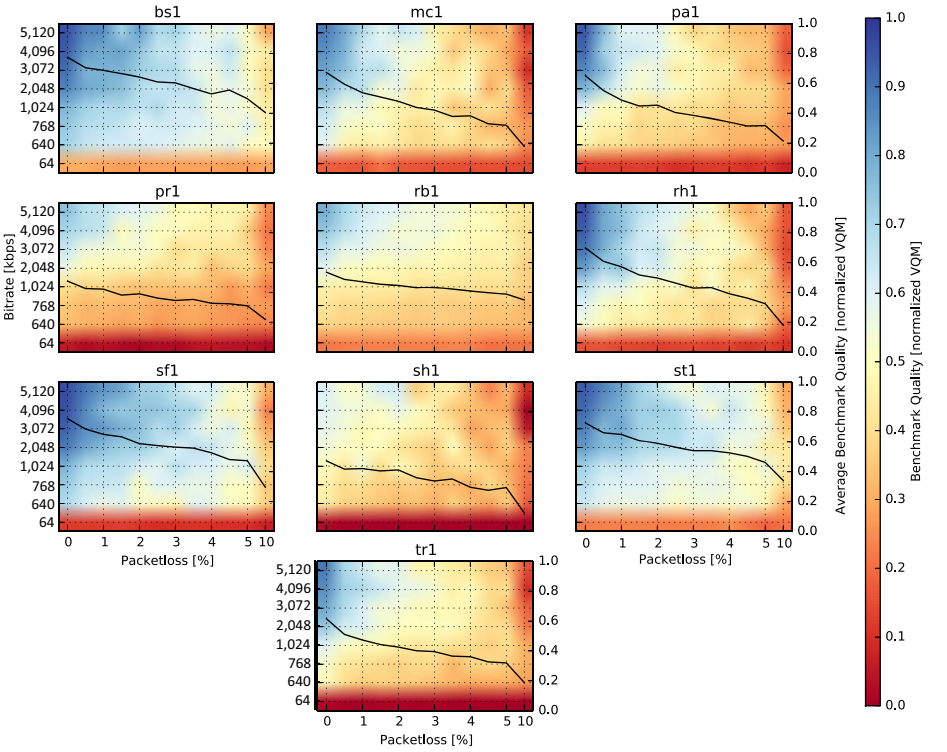


**Figure 5.**
Benchmark quality of the 960 videos of the impaired video set

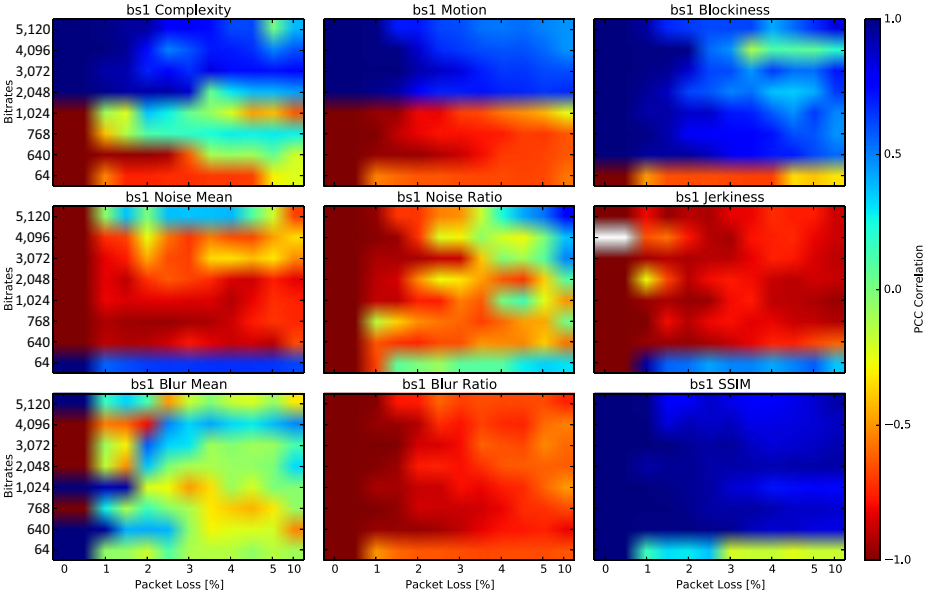


**Figure 6.**
NR metrics and SSIM correlation to the benchmark quality for selected video type bs1 and all impairment conditions

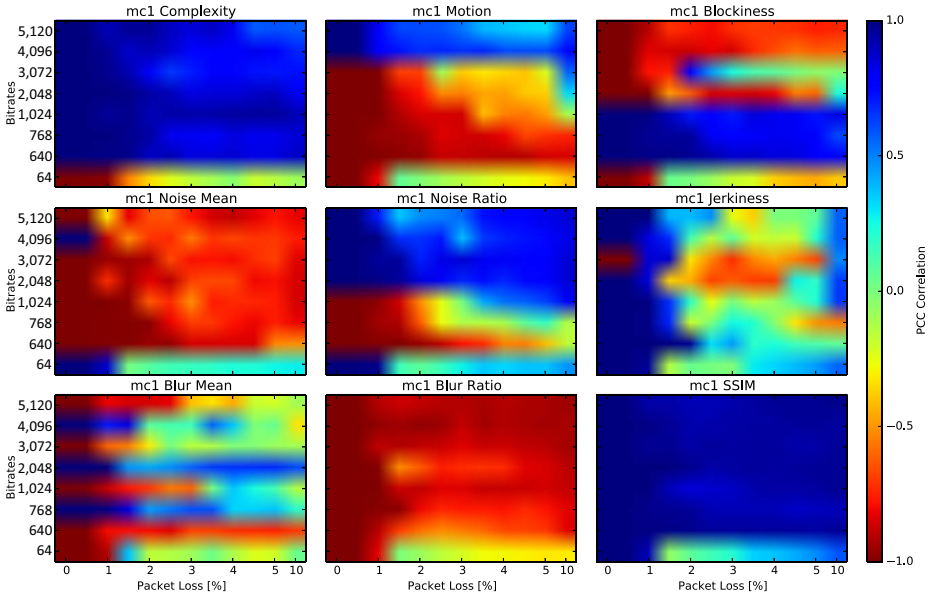


**Figure 7.**
NR metrics and SSIM correlation to the benchmark quality for selected video type mc1 and all impairment conditions

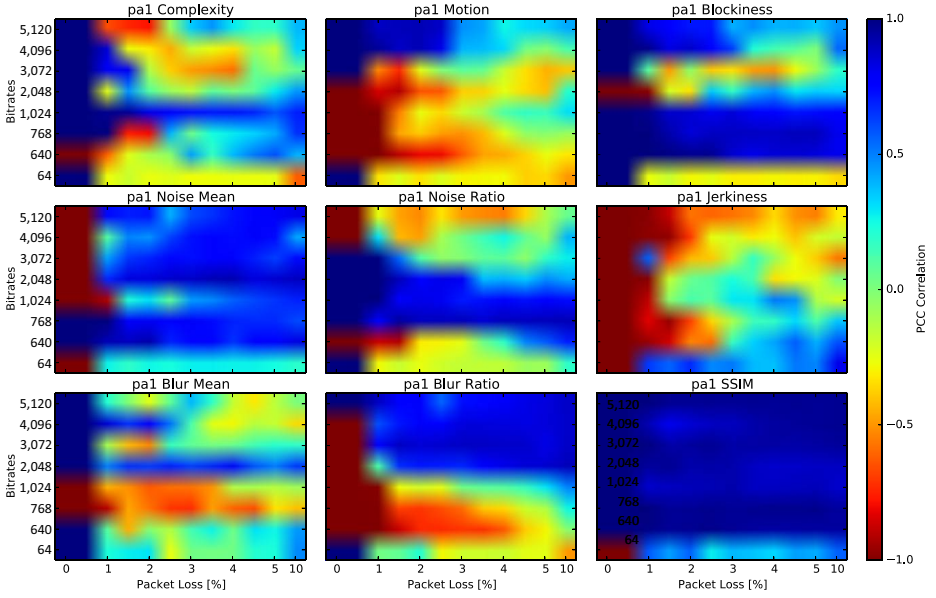


**Figure 8.**
NR metrics and SSIM correlation to the benchmark quality for selected video type pa1 and all impairment conditions

is found at the lower levels of packet loss and high bitrates. The color gradually turns to yellow as the network packet loss increases. In none of the cases, quality degrades beyond values of 0.1-0.3 (red-dark orange). One interesting finding is that as the bitrate decreases, the influence of packet loss is less noticeable. In all the tested cases, bitrates of

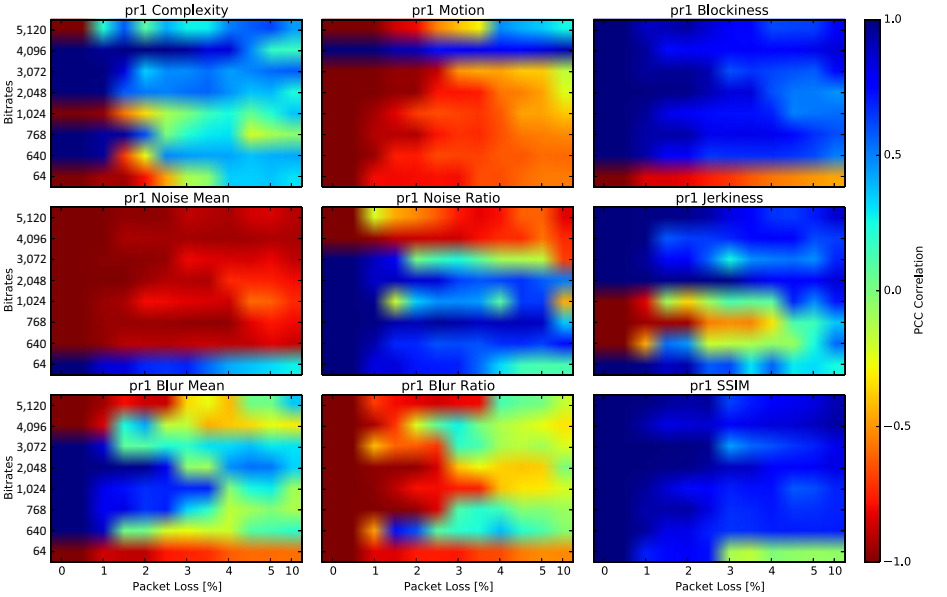


**Figure 9.**
NR metrics and SSIM correlation to the benchmark quality for selected video type pr1 and all impairment conditions

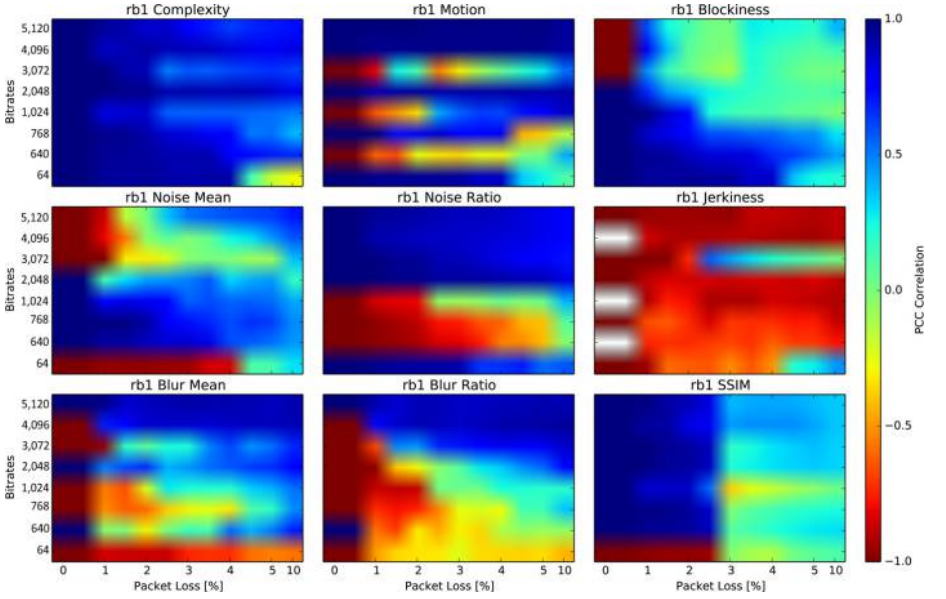


**Figure 10.**
NR metrics and SSIM correlation to the benchmark quality for selected video type rb1 and all impairment conditions

64 kbps and 640 kbps get hardly impaired by the network conditions (constant color between yellow and orange depending on the video). Furthermore, as it was hinted in the study of the original video set, the video types are affected by packet loss to a greater or a lower extent, depending on their dynamic composition. For example, the videos Park

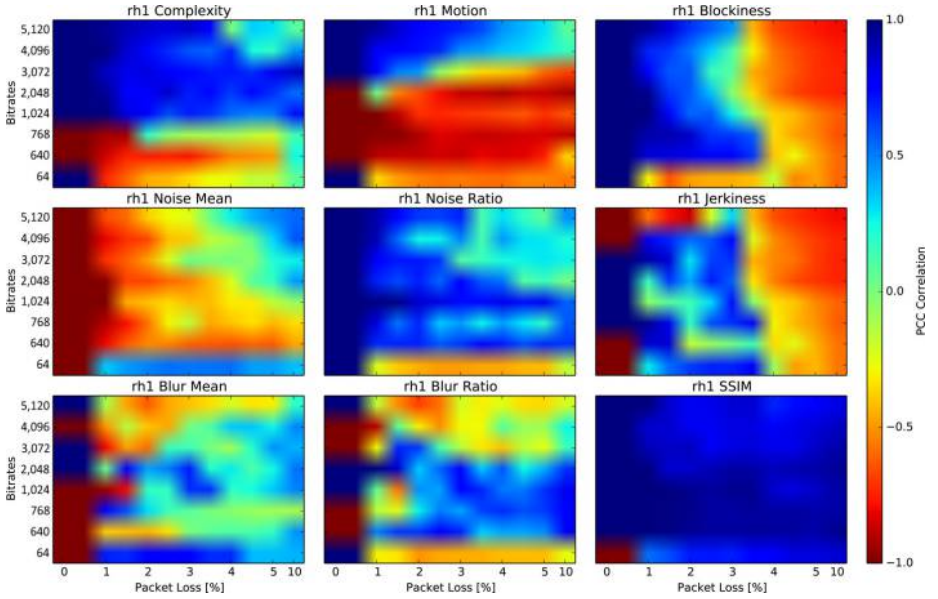


**Figure 11.**
NR metrics and SSIM correlation to the benchmark quality for selected video type rh1 and all impairment conditions

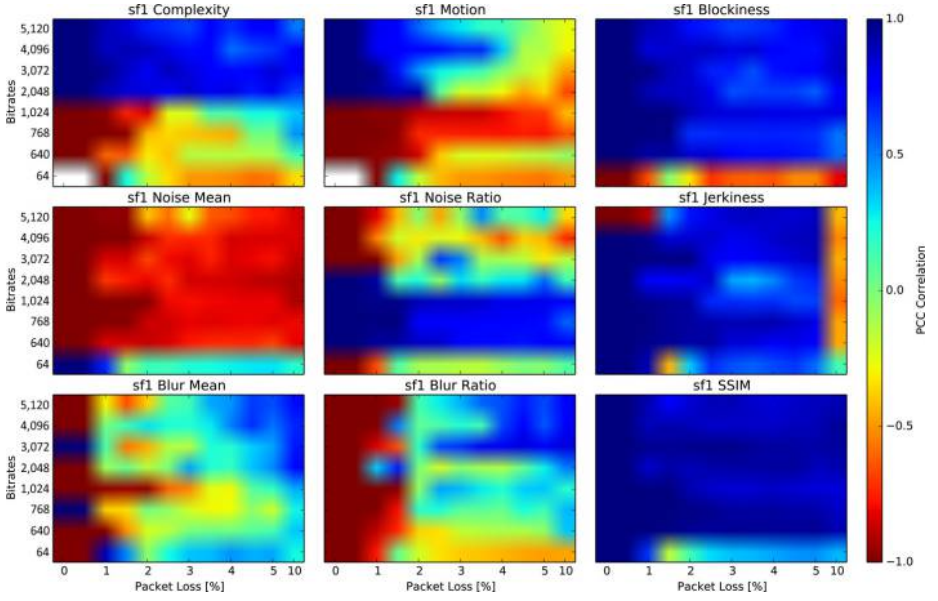


**Figure 12.**
NR metrics and SSIM correlation to the benchmark quality for selected video type sf1 and all impairment conditions

run (pr1) or River bed (rb1) suffer more degradation from the compression than from the influence of packet loss (near constant colors in all the bitrates). In the other extreme, cases such as the Mobile calendar (mc1), Blue sky (bs1) or Rush hour (rh1) suffer great degradation from packet loss. Finally, videos like the Shields (sh1) or the Tractor (tr1)

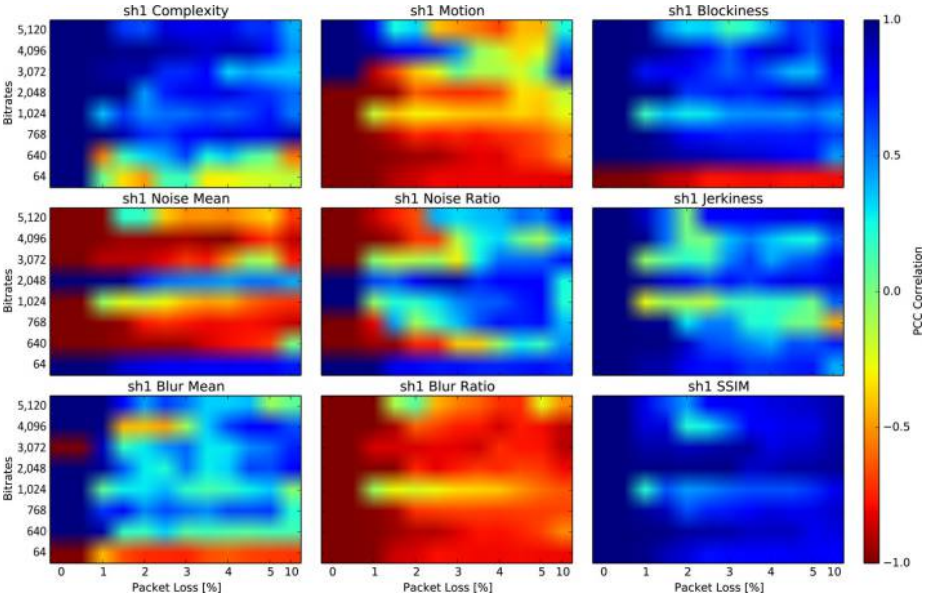


**Figure 13.**
NR metrics and SSIM correlation to the benchmark quality for selected video type sh1 and all impairment conditions

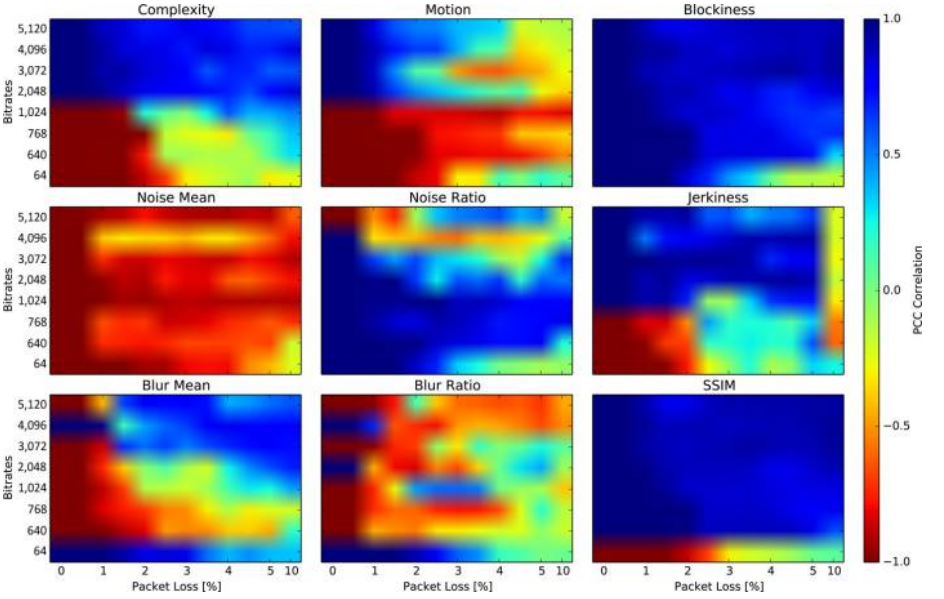


**Figure 14.**
NR metrics and SSIM correlation to the benchmark quality for selected video type st1 and all impairment conditions

show a counter-intuitive behavior because they suffer faster degradation for low packet loss (until roughly 3 per cent) and slow degradation from that point on.

Figures 6-15 show the correlation results for the ten video types of our data set. Each figure presents a different video type, whereby each colormap shows the correlation of

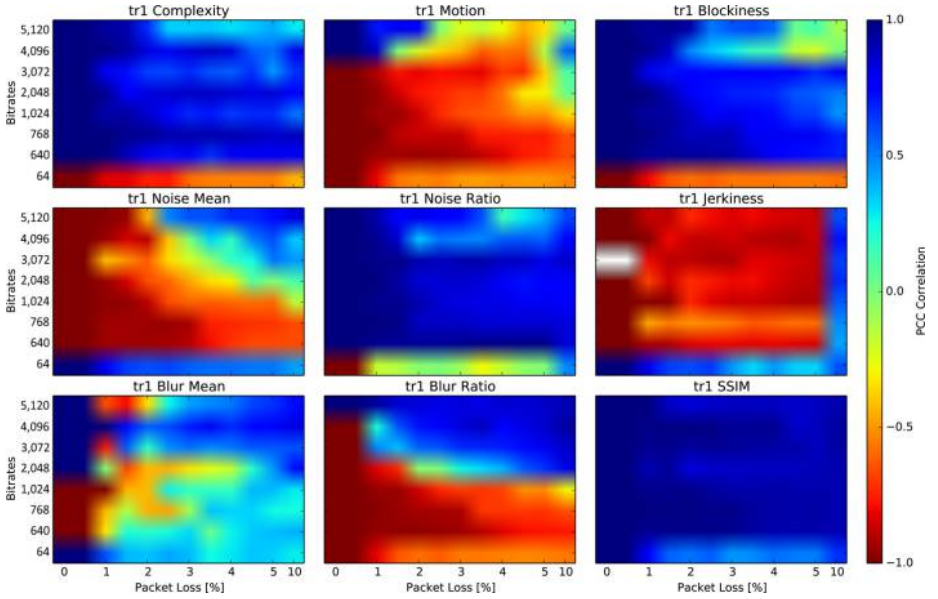


**Figure 15.**
NR metrics and SSIM correlation to the benchmark quality for selected video type tr1 and all impairment conditions

a specific metric to the benchmark quality. The *x*-axis presents the packet loss level, and the *y*-axis shows the correlations for each of the bitrates under scrutiny. As in the previous section, a dark blue value denotes full correlation (quality = 1). As the correlation degrades, the color goes from light blue to green (0 or no correlation), yellow and final red (−1, the two metrics are anti-correlated, that is, they follow opposite trends). Some of the cases in which an anti-correlation takes place (red) present, at certain levels of packet loss, an apparent improvement in correlation. This is just an effect of the metric failing to perform an accurate measurement due to high losses (i.e. missing frames).

As in the previous part of the analysis, we can first take a look at the bitstream features, complexity and motion (first and second plots of the first row, in each of the ten plots). In general, it can be seen that the scene complexity correlates better with the benchmark quality as the bitrate increases. The correlation occurs up to a certain level of packet loss, changing across different types of video and bitrate but ranging between 2.5 and 4 per cent, from which the correlation starts degrading. Depending on the video type, the lower bitrates present low correlation or even anti-correlation. For example, while in the videos River bed (rb1, Figure 10) and Shields (sh1, Figure 13) the complexity presents high levels of correlation for all the bitrates, the remaining eight videos present well-defined bitrate thresholds from which the correlation starts and below which the metric is completely anti-correlated with the benchmark quality. This threshold is, for example, 2 Mbps for the Blue sky (bs1, Figure 6) or 640 kbps for the Mobile calendar (mc1, Figure 7). An extreme case is the one presented in the video Pedestrian area (pa1, Figure 8). In this case, the correlation occurs not only depending on the level of packet loss but also on the compression bitrate. In this way, no correlation is present for compression rates of 5 and 1 Mbps, but it appears for all the other rates. The video motion follows a similar behavioral pattern as for the complexity. Correlation occurs

predominantly on high-quality videos and, as the packet loss level increases, the correlation slowly decreases (lighter blue). However, in contrast to what was found in the complexity study, the correlation of the motion occurs in a more reduced number of places, and no video type presents a motion correlation in any of its variants and network conditions. One extreme example is the River bed (rb1, Figure 10), in which, depending on the bitrate, the motion either fully correlates or anti-correlates.

The blockiness (third plot, first row of the ten figures) presents, in general, higher correlations for low bitrates (up to 1-2 Mbps depending on the video) and low level of packet losses (up to 2-4 per cent). For the high bitrates compression variants, the trend depends fully on the video type. In seven out of the ten video types, correlations are to be found also at high bitrates up to a certain level of packet loss. From the remaining three video cases, Mobile calendar (mc1, Figure 7) and the River bed (rb1, Figure 10) present no blockiness correlation at high bitrates, and the Pedestrian area (pa1, Figure 8) presents correlation dependent on the bitrate. While for the 5 and 4 Mbps variants of the video, the blockiness is correlated up to close to 5 per cent packet loss level, in the 2 and 1 Mbps variant, the correlation is negative.

If we now focus on the components of the image NR-metric developed by Choi *et al.* (2009), noise and blur (first and second plots, second and third row of the 10 Figures), then we can see that the correlation is poor for most of the video types but completely dependent on video type and bitrate. For example, while the noise mean fully anti-correlates in most of the cases, it correlates for certain bitrate variants of the Pedestrian area (pa1, Figure 8) and Shields videos (sh1, Figure 13).

The jerkiness' (third plot, second row for the 10 Figures) correlation depends completely on the video type and bitrate. On the one hand, videos such as the Blue sky (bs1, Figure 6), the Pedestrian area (pa1, Figure 8) or the River bed (rb1, Figure 10) present a complete lack of correlation between jerkiness and quality. On the other hand, the Park run (pr1, Figure 9) and Sunflower (sf1, Figure 12) show good correlation between the metric and the quality. Finally, others, such as the Station (st1, Figure 14), present a combined pattern, in which correlation appears at high bitrates and a complete lack of it at low bitrates and high packet losses. The different behaviors come from the composition of the videos. Furthermore, it is worth noticing that the nature of the transmissions (the real time protocol) avoids the appearances of freezes or time laps, while it is more vulnerable to blocks and bitrate reductions.

Finally, SSIM (third plot, third row for the ten figures) correlates in nearly all the cases and bitrates with the exception of very low bitrates (64 kbps) and very high levels of packet loss (roughly 2-4 per cent). One clear example of this is the case of the River bed (rb1, Figure 10), in which for all bitrate compression (except 64 Kbps), the correlation is nearly perfect until 3 per cent packet loss to decay to zero correlation for higher packet loss levels. This general correlation proves that even if considered a lesser FR metric in terms of accuracy, SSIM is still a valid indicator of the degradation of video services.

From these results, we can conclude that for each metric, we can always find a specific range of good operational quality. However, none of the NR metrics operates uniformly well, that is, under an overall representative range of conditions. This result suggests that contrary to what is normally done in the objective QoE assessment, NR metrics cannot be deemed appropriate to evaluate network-impaired video streams. Furthermore, our work confirms that FR metrics such as VQM are indeed accurate not

only to assess quality loss due to compression (as it is well known) but also to assess quality loss in the presence of substantial packet and frame losses.

## 6. Conclusions

In this work, we wanted to find out the extent by which simple NR metrics would work for network-impaired video streams. This was motivated by some of our own earlier studies on QoE evaluation of video streaming systems, which unveiled substantial issues with the current methods particularly under lossy conditions (Mocanu *et al.*, 2014a, 2014b, 2014c; Torres Vega *et al.*, 2014). The contribution of this paper is, first, to verify that VQM is reliable even at high packet loss rates. Our stress tests cover a broad range of video types and network impairments up to 10 per cent packet loss rates, showing that despite being computationally intensive, VQM can be used as a reliable benchmark to evaluate other lightweight metrics. We could then carry out a comparative experimental survey covering a range of NR metrics, including the lightweight FR metric SSIM, establishing that, indeed, NR-metrics are accurate only within restricted operational conditions. None of the tested metrics operated accurately on a sufficiently broad set of test cases. This leads to the conclusion that common practices on NR QoE assessment studies should be considerably revisited, particularly when evaluating real-time streams over realistic network conditions.

On a more positive note, we can see that if combined, the different NR metrics could actually cover a broad range of operational conditions. Armed with these results, our next move is to study new methods to automatically determine the operational range of each metric and to build a new hybrid metric that combines the simplicity of NR algorithms with the accuracy of VQM.

## References

Borer, S. (2010), "A model of jerkiness for temporal impairments in video transmission", *Proceeding of the Second International Workshop on Quality of Media Experience (QoMEX), Trondheim*.

Brandão, T. and Queluz, M. (2010), "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.

Chikkerur, S., Sundaram, V., Reisslein, M. and Karam, L.J. (2011), "Objective video quality assessment methods: a classification, review, and performance comparison", *TBC*, Vol. 57 No. 2, pp. 165-182.

Choi, M.G., Jung, J.H. and Jeon, J.W. (2009), "No-reference image quality assessment using blur and noise", *International Journal of Electrical, Computer, Energetic, Electronic and Communication Engineering*, Vol. 3 No. 2.

Ciancio, A.G., da Costa, A.L.N.T., da Silva, E.A.B., Said, A., Samadani, R. and Obrador, P. (2011), "No-reference blur assessment of digital pictures based on multifeature classifiers", *IEEE Transactions on Image Processing*, Vol. 20 No. 1, pp. 64-75.

Ferzli, R. and Karam, L.J. (2006), "Human visual system based no-reference objective image sharpness metric", *2006 IEEE International Conference on Image Processing, Atlanta, GA*, pp. 2949-2952.

Hemami, S.S. and Reibman, A.R. (2010), "No-reference image and video quality estimation: applications and human-motivated design", *Signal Processing: Image Communication*, Vol. 25 No. 7, pp. 469-481.

Hu, J. and Wildfeuer, H. (2009), "Use of content complexity factors in video over ip quality monitoring", *International Workshop on Quality of Multimedia Experience*, QoMEx, San Diego, CA, pp. 216-221.

Keimel, C., Habigt, J. and Diepold, K. (2012), "Hybrid no-reference video quality metric based on multiway plsr", in *'EUSIPCO 2012: 20th European Signal Processing Conference'*, *Bucharest*, pp. 1244-1248.

Kendall, M.G., Stuart, A. and Ord, J.K. (Eds) (1987), *Kendall's Advanced Theory of Statistics*, Oxford University Press, New York, NY.

Le Callet, P., Moeller, S. and Perkis, A. (Eds) (2012), "Qualinet white paper on definitions of quality of experience", *COST Action IC 1003*, Laussanne.

Liotta, A. (2013), "The cognitive NET is coming", *IEEE Spectrum*, Vol. 50 No. 8, pp. 26-31.

Liotta, A., Mocanu, D.C., Menkovski, V., Cagnetta, L. and Exarchakos, G. (2013), "Instantaneous video quality assessment for lightweight devices", *Proceeding of International Conference on Advances in Mobile Computing, MoMM'13*, *New York, NY*, pp. 525-531.

Liu, H. and Heynderickx, I. (2008), "A no-reference perceptual blockiness metric", *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, *Las Vegas, NV*.

Menkovski, V., Exarchakos, G. and Liotta, A. (2011), "The value of relative quality in video delivery", *Journal of Mobile Multimedia*, Vol. 7 No. 3, pp. 151-162.

Menkovski, V., Exarchakos, G., Liotta, A. and Sanchez, A.C. (2010), "Quality of experience models for multimedia streaming", *IJMCMC*, Vol. 2 No. 4, pp. 1-20.

Mocanu, D., Exarchakos, G. and Liotta, A. (2014b), "Deep learning for objective quality assessment of 3d images", *IEEE International Conference on Image Processing (ICIP)*, *Paris*, pp. 758-762.

Mocanu, D.C., Santandrea, G., Cerroni, W., Callegati, F. and Liotta, A. (2014a), "Network performance assessment with quality of experience benchmarks", *International Conference on Network and Service Management (CNSM)*, Rio de Janeiro.

Mocanu, D., Liotta, A., Ricci, A., Vega, M. and Exarchakos, G. (2014c), "When does lower bitrate give higher quality in modern video services?", *IEEE Network Operations and Management Symposium (NOMS)*, Krakow, pp. 1-5.

Mocanu, D.C., Exarchakos, G., Bou-Ammar, H. and Liotta, A. (2015), "Reduced reference image quality assessment via boltzmann machines", *IM in IFIP/IEEE International Symposium on Integrated Network Management*, Ottawa, ON, 11-15 May, pp. 1278-1281.

Perra, C. (2014), "A low computational complexity blockiness estimation based on spatial analysis", *IEEE 22nd Telecommunications Forum*, Belgrade.

Pinson, M.H. and Wolf, S. (2004), "A new standardized method for objectively measuring video quality", *IEEE Transactions on Broadcasting*, Vol. 50 No. 3, pp. 312-322.

Rank, K., Lendl, M. and Unbehauen, R. (1999), "Estimation of image noise variance", *IEEE Proceedings Visual, Image, Signal Processing*, Vol. 146 No. 2, pp. 80-84.

Reibman, A., Sen, S. and der Merwe, J. (2005), "Analyzing the spatial quality of internet streaming video", *Proceedings of International Workshop on Video Processing and Quality Metrics for Consumer Electronics*, Las Vegas.

Seshadrinathan, K., Soundararajan, R., Bovik, A.C. and Cormack, L.K. (2010a), "Study of subjective and objective quality assessment of video", *IEEE Transactions on Image Processing*, Vol. 19 No. 6, pp. 1427-1441.

Seshadrinathan, K., Soundararajan, R., Bovik, A.C. and Cormack, L.K. (2010b), "A subjective study to evaluate video quality assessment algorithms", in Rogowitz, B.E. and Pappas, T.N. (Eds), *SPIE Proceedings on Human Vision and Electronic Imaging*, Vol. 7527, p. 75270.

Shahid, M., Rossholm, B. and Lövström, B. (2011), "A reduced complexity no-reference artificial neural network based video quality predictor", *Proceeding of the International Congress on Image and Signal Processing*, Shanghai.

Shahid, M., Rossholm, A., Lövström, B. and Zepernick, H. (2014), "No-reference image and video quality assessment: a classification and review of recent approaches", *EURASIP Journal on Image and Video Processing*, Vol. 40 No. 1.

Shanableh, T. (2011), "Prediction of structural similarity index of compressed video at a macroblock level", *IEEE Signal Processing Letters*, Vol. 18 No. 5.

Staehle, B., Binzenhöfer, A., Schlosser, D. and Boder, B. (2008), "Quantifying the influence of network conditions on the service quality experienced by a thin client user", *14. GI/ ITG Konferenz Messung, Modellierung und Bewertung von Rechen- und Kommunikationssystemen (MMB 2008)*, Dortmund.

Suárez, F.J., García, A., Granda, J.C., Garcia, D.F. and Nuño, P. (2015), "Assessing the qoe in video services over lossy networks", *Journal of Network and Systems Management*, Vol. 24 No. 1, pp. 116-139.

Torres Vega, M., Giordano, E., Mocanu, D.C. and Liotta, A. (2015a), "Cognitive no-reference video quality assessment for mobile streaming services", *Proceeding of the 7th International Workshop on Quality of Multimedia Experience (QoMex)*, Pylos-Nestoras.

Torres Vega, M., Mocanu, D.C., Barresi, R., Fortino, G. and Liotta, A. (2015b), "Cognitive streaming on android devices", *Proceeding of the 1st. IEEE/IFIP IM 2015 International Workshop on Cognitive Network & Service Management*, Ottawa, ON.

Torres Vega, M., Sguazzo, V., Mocanu, D.C. and Liotta, A. (2015c), "Accuracy of no-reference quality metrics in network-impaired video streams", *13th International Conference on Advances in Mobile Computing and Multimedia (MoMM2015)*, Vol. 12 No. 1.

Torres Vega, M., Zou, S., Mocanu, D.C., Tangdiongga, E., Koonen, A.M.J. and Liotta, A. (2014), "End-to-end performance evaluation in high-speed wireless networks", *The International Conference on Network and Service Management (CNSM)*, Rio de Janeiro.

Wang, Z., Lu, L. and Bovik, A.C. (2004), "Video quality assessment based on structural distortion measurement", *Signal Processing: Image Communication*, Vol. 19 No. 2, pp. 121-132.

Winkler, S. (2005), *Digital Video Quality: Vision Models and Metrics*, 1 edn, Wiley.

Wu, H.R. and Yuen, M. (1997), "A generalized block-edge impairment metric for video coding", *IEEE Signal Processing Letters*, Vol. 4 No. 11, pp. 317-320.

Zinner, T., Hohlfeld, O., Abboud, O. and Hossfeld, T. (2010), "Impact of frame rate and resolution on objective QoE metrics", *International Workshop on Quality of Multimedia Experience, QoMEx 2009*, Trondheim.

## About the authors

Maria Torres Vega received her MSc degree in Telecommunication Engineering from the Polytechnic University of Madrid, Spain, in 2009. Between 2009 and 2013, she worked as a Software and Test Engineer in Germany with focus on Embedded Systems and Signal Processing. In October 2013, she decided to go back to academia, and since then, she is a PhD Student at the Eindhoven University of Technology. Her research interests include, but are not limited to, computer vision, quality of service and quality of experience in multimedia systems, autonomic

management of wireless networks, artificial intelligence and machine learning. Maria Torres Vega is the corresponding author and can be contacted at: m.torres.vega@tue.nl

Vittorio Sguazzo is currently finishing his Bachelor's degree in Telecommunication Applications in the faculty of computer engineering at Salerno University, Italy. This work was performed during his internship at the Smart Networks group of the Eindhoven University of Technology.

Decebal Constantin Mocanu received the BEng degree in Computer Science from Polytechnic University of Bucharest, Romania, in 2010 and the MSc degree in Artificial Intelligence from Maastricht University, The Netherlands, in 2013, for which he received the best "Master AI Thesis Award". In parallel with his bachelor and master studies, from 2001 until 2013, he worked as a Software Engineer. Starting from September 2013, he is a PhD Student at Eindhoven University of Technology, The Netherlands. His research interests include, but are not limited to, artificial intelligence, machine learning, complex networks, communication networks and computer vision.

Antonio Liotta holds the Chair of Communication Network Protocols at the Eindhoven University of Technology (NL), where he leads the Smart Networks team. Liotta is the Editor-in-Chief of the book series *Internet of Things: Technology, Communications and Computing* (Springer) and an Associate Editor of the *Journal of Network and System Management* (Springer) and of the *International Journal of Network Management* (Wiley). He investigates topical issues in the area of computer and multimedia networking, with emphasis on cognitive systems for wireless and sensor networks. He is the author of *Networks for Pervasive Services: six ways to upgrade the Internet* (Springer). His work has recently been featured in IEEE Spectrum in *The cognitive Net is coming*.