EURASIP Journal on Image and Video Processing
a SpringerOpen Journal

RESEARCH   **Open Access**

Comparing apples and oranges: assessment of the relative video quality in the presence of different types of distortions

Ulrich Reiter[1*], Jari Korhonen[2] and Junyong You[1]

Abstract

Video quality assessment is essential for the performance analysis of visual communication applications. Objective metrics can be used for estimating the relative quality differences, but they typically give reliable results only if the compared videos contain similar types of quality distortion. However, video compression typically produces different kinds of visual artifacts than transmission errors. In this article, we focus on a novel subjective quality assessment method that is suitable for comparing different types of quality distortions. The proposed method has been used to evaluate how well different objective quality metrics estimate the relative subjective quality levels for content with different types of quality distortions. Our conclusion is that none of the studied objective metrics works reliably for assessing the co-impact of compression artifacts and transmission errors on the subjective quality. Nevertheless, we have observed that the objective metrics' tendency to either over- or underestimate the perceived impact of transmission errors has a high correlation with the spatial and temporal activity levels of the content. Therefore, our results can be useful for improving the performance of objective metrics in the presence of both source and channel distortions.

Keywords: multimedia communication, video quality assessment

1. Introduction

In video streaming applications, visual quality is often affected by two generic, but fundamentally different types of quality distortion: source distortion and channel distortion. Source distortion is derived from video compression that is necessary to comply with the bandwidth limitations of the communication system. Channel distortion is caused by transmission errors (packet losses and/or bit errors), occurring in the communication channel. In practice, source distortion can often be decreased at the cost of increased channel distortion, and vice versa. This is because higher quality requires higher bitrates, which in turn leaves a smaller proportion of the channel capacity to be allocated for error correction via redundancy (forward error correction–FEC) or retransmission. Respectively, decreasing the bitrate of the compressed video bitstream

increases source distortion, but allows more bandwidth to be used for protection against channel distortion [1].

The artifacts caused by transmission errors are qualitatively very different from compression artifacts. Video compression affects the overall quality of the video, whereas transmission errors typically appear in spatially and temporally limited areas. Unfortunately, the established objective quality metrics have typically not been cross-validated sufficiently well with different distortion types, and the primary focus of objective quality assessment has traditionally been on compression artifacts, and only a few objective metrics consider transmission errors or other types of distortion, e.g., packet loss [2]. For example, traditional peak signal-to-noise ratio (PSNR) is well known to be overly sensitive to certain artifacts, such as contrast changes and spatial shift [3]. In addition, PSNR results are only meaningful for comparison of distorted sequences showing the same content [4]. Even the more advanced metrics have similar limitations for their scope of use [5]. This is why accurate quality estimates for video sequences cannot be expected from a metric optimized for

* Correspondence: reiter@q2s.ntnu.no
[1]Centre for Quantifiable Quality of Service in Communication Systems (Q2S), Norwegian University of Science and Technology, O.S. Bragstads plass 2E, 7491 Trondheim, Norway
Full list of author information is available at the end of the article

Springer

compression artifacts when channel distortion is involved, and vice versa. Unfortunately, subjective quality evaluation requires quite a lot of human resources and time for preparation. In a typical subjective assessment study, the reference video sequence and the impaired test sequence are shown in parallel and test subjects give their subjective rating on the quality of the test sequence, compared to [6]. However, comparing different types of distortions and mixtures of them can easily lead to an enormous set of test cases. In order to reduce the number of required test cases and obtain more accurate results, we have proposed a novel method for a comparative subjective video quality assessment, termed double-stimulus adjustable quality fixed anchor (DSAQFA) [7]. In our method, two video sequences are shown in parallel, but instead of rating the difference, the user is asked to adjust the quality of the scalable sequence to match with the fixed sequence. In this way, it is possible to compare the subjective quality distortions of different types with a reasonable number of trials.

In our earlier studies, we have used the method successfully to evaluate the subjective quality differences between video sequences with source and channel distortion. The results showed that the perceptual impact of channel distortion may be either overestimated or underestimated when PSNR is used as a quality metric, depending on the content of the video sequence [7,8]. In this article, we extend our analysis beyond PSNR. For this purpose, we have used several well-established objective video quality metrics, each of which has been reported to outperform PSNR. The fundamental question is then, how well are these different objective metrics capable of predicting the relative perceived quality levels when video sequences with qualitatively different types of distortions are compared?

The remainder of this article is organized as follows. In Section 2, we explain the background and define the problem more thoroughly. In Section 3, we explain our test methodology and the practical experiments. In Section 4, we analyze the results obtained in our experiments. A discussion of the results is included in Section 5, and finally, some concluding remarks are given in Section 6.

2. Background

Owing to the rapid advances in mobile communications and the raise of social online communities, multimedia traffic in wireless networks is growing rapidly. On the other hand, radio spectrum is a scarce resource, and it is getting more and more important to use the wireless bandwidth efficiently [9]. In order to provide the users with technology that guarantees the best possible quality of experience, it is essential to measure the impact of all relevant factors on quality accurately and efficiently. In the context of wireless networking, compression and transmission errors are with no doubt among those factors. In this section, we explain the relevant fundamentals of such a networking scenario employing wireless video transmission, as well as the video quality assessment methods relevant to our study.

2.1. Networking application scenario

In a typical video streaming or teleconferencing application, some kind of rate control mechanism is involved to cope with congestion in the network. Usually, rate control algorithms use packet losses as indicator for congestion. When a receiver detects one or more packet losses, it sends a feedback message to the sender to request a lower transmission rate, which is supposed to relieve congestion. Due to the real-time nature of the communication, the data transmission rate should match the source coding rate. Small variations along time are acceptable, since they can be compensated with buffering. In practice, there are several methods for adjusting the source coding rate. If live content is concerned, then encoding parameters of the compression algorithm can be changed on-the-fly to obtain the target bitrate. For prerecorded content, the simplest method is *bitstream switching*: several versions of the content are stored at the server, each compressed at a different quality level to produce versions with different bitrates. The sender chooses the version of highest quality that fits into the available (limited) bandwidth. A more advanced method is to use *layered coding*; here, an encoded bitstream is divided into layers, and different quality levels can be obtained by adding or removing parts of the bitstream (layers).

Conventional rate control is basically the only way to combat congestion, when traditional wired networks are concerned. However, the rapid advancement of wireless networking has challenged this wisdom. In a wireless radio channel, physical transmission errors are far more common than in fixed cables, and practically all real-life wireless networking protocols perform some kind of error control to recover bit errors, usually by retransmitting the erroneous packets. Obviously, retransmissions consume part of the channel capacity, which may lead to less capacity left for the original stream. This is where a fundamental question arises: in order to achieve the highest possible *perceived quality*, would it be more beneficial to keep a higher transmission rate and allow some transmission errors, either packet losses or bit errors in the content, or is it better to reduce the source bitrate to obtain an error-free transmission [10]? Assuming that the first alternative is chosen, it would be possible to switch off retransmissions and pass on damaged packets to the application instead of requesting a retransmission. However, the question is impossible to answer, if we cannot reliably compare the subjective

quality degradations caused by channel and source distortions, respectively.

Several studies have been published regarding the question if it is a good idea to deliver erroneous packets to the application. One possibility to implement such functionality is to use a protocol like UDP Lite [11], employing a partial checksum for bit error detection. Partial checksums cover only the most vulnerable parts of the packet, such as protocol headers. It has widely been accepted that in the presence of bit errors, UDP Lite can significantly improve the throughput [1,11-14]. However, the improved usage of channel capacity comes at the cost of bit errors appearing in the coded content. Therefore, the benefits of using UDP Lite depend highly on how well the bit errors are handled at the application layer. A wide range of different partial error protection schemes have been proposed for error prone transmission. According to Singh et al. [12], some video quality improvements can be gained with UDP Lite together with an error resilient codec, but the results are highly dependent on the bit error characteristics. Other researchers, such as Khayam et al. [13] and Masala et al. [14], have proposed different FEC and partitioning strategies in adjunction with UDP Lite to obtain better performance. In our earlier study, we have observed relatively large improvements in terms of PSNR, when a UDP Lite approach is compared to conventional rate control in a congested radio channel [1].

One limitation for the majority of proposals and studies related to UDP Lite and similar schemes is that the quality comparisons have been made by computing PSNR [1,13,14] or analyzing the network parameters (delay, burst length of affected frames, etc.) [12]. As we have discussed in Section 1, these approaches can only give very rough estimates of the quality as perceived by a human. This is the major motivation for studying the relative perceptual impact of source and channel distortion in video sequences.

2.2. Video quality assessment

Even though significant efforts have been invested to develop objective models for measuring video quality, the scope of use for even the best performing objective quality metrics is still rather limited, and reliable results should not be expected from models that have not been verified for the particular use case in question [5]. This is why subjective quality assessment is still required in many situations, not least in performance evaluation of objective metrics. In a typical subjective quality assessment study, test subjects are asked to rate the test video using a given quality scale. The average of the scores, mean opinion score (MOS), then represents the subjective quality of the test video. However, different subjects often interpret verbal descriptors differently, especially when taking into account that the terms usually are

translated into different languages for subjects from different countries. A similar problem applies to the intervals between the quality scale labels (i.e., distance between 'good' and 'fair' is supposed to be equal to the one between 'poor' and 'bad'). A rating task is easier for the test subjects if the original sequence is available for comparison, and this is why double stimulus methods are often preferred in subjective assessment studies.

To minimize the impact of random environmental factors on the results, there are standards defining the arrangements for subjective quality evaluations. ITU-R BT.500-11 [15] describes a number of approaches for the subjective assessment of television picture quality. Among them, double stimulus impairment scale (DSIS) and double stimulus comparison scale (DSCS) methods are the most relevant for this study. In DSIS, the task of the test subjects is to evaluate the impairment of the test sequence with respect to the reference sequence, using a 5-point scale from "very annoying" to "imperceptible". In turn, DSCS is a more suitable method if the test sequence may be of higher perceived quality than the reference sequence, since it uses a comparative scale ranging from "much worse" to "much better". Unfortunately, the general problems with vocabulary and intervals between the quality scale labels remain unsolved with these methods [16]. Another problem is that these recommendations were developed for television systems. For today's applications based on wireless networks and mobile devices, as for example multimedia conferencing applications, the vocabulary used for the quality scale is unsuitable, and subjects' responses can be expected to be biased toward the bottom of the scale [16].

In the DSAQFA method proposed in our earlier study [7], test subjects are instructed to select from a range of adjustable quality sequences the one that matches as closely as possible the subjective quality of a fixed reference sequence. The proposed method is especially well suited for comparing video sequences containing different types of distortions, such as source and channel distortions. Since there is no need for a rating on a scale, problems with vocabulary and different interpretations of the scale level denominators can be avoided. In our experience, the proposed method also saves time, since there is less training needed for the test subjects. Also, in many scenarios, the number of presentations that are required can be reduced compared to rating scale-based approaches, since the process allows subjects to check a wide range of degrees of distortion within the same trial. The details of the method are explained in Section 3.1.

3. Methodology

Most of the related work in the field of subjective video quality assessment uses different rating scales, either absolute MOS scales or comparative scales, as in DSIS

and DSCS methods. In this section, the DSAQFA method designed to overcome the limitations of rating scales is introduced in detail, followed by a description of practical experiments in which the method has been employed successfully [7,8].

3.1. DSAQFA method

In order to compare different types of quality distortions, it seems appropriate to use a test methodology that is independent of vocabulary and scales. The DSAQFA method builds on explicitly ignoring quality- or impairment-scales, thus avoiding the error-prone process of transforming one scale (suitable for the evaluation of one type of quality distortion) into another one (suitable for a different type of quality distortion). Instead, we have proposed a methodology that simply requires test subjects to adjust the quality of one stimulus to match the perceived quality of another fixed stimulus.

In DSAQFA comparisons, subjects are presented with a fixed stimulus of given impairments of one or more types, next to an adjustable test stimulus with different type (or combination) of impairment. Subjects are then asked to adjust the perceived quality of the test stimulus such that it matches best with the perceived quality of the fixed stimulus. As the type of impairment in the two is different, the match can never be perfect, even if the adjustable quality was to be controlled continuously (i.e., not stepwise)–hence, the subject is required to compare apples and oranges. In spite of the concerns that such a comparison raises–after all, subjects may find it particularly difficult to integrate a number of different artifacts (as opposed to rating one quality attribute at a time)–we have found this methodology to be requiring a minimum amount of training, usually resulting in assessment sessions of shorter duration.

The test software used for DSAQFA experiments has been described in [7]. The functionality of the software on a generic level is illustrated in Figure 1. Quality adjustment has been implemented by using several raw video source files with the same content, but different quality levels. The main element of the user interface is a slider that is used to choose the video source file among n sequences with different quality level, together forming an adjustable video sequence. In our experiments, the video sequences were rather short (around 10 s each) and they were automatically repeated from the beginning when the end of the file was reached. When the subject pressed the "ok" button, the slider position representing the current quality of the adjustable video was stored in a log file and the next test case was started.

In the analysis phase, an objective quality metric can be applied to the fixed sequence and all the sequences forming the adjustable video, to obtain quality indices for each sequence. When a sufficient number of test subjects have performed the test, the average quality of the sequences chosen by the subjects is supposed to describe the quality level of the adjustable sequence that is perceptually equivalent to the quality level of the fixed sequence. In an ideal case, the quality index of the fixed sequence and the average quality index of the voted
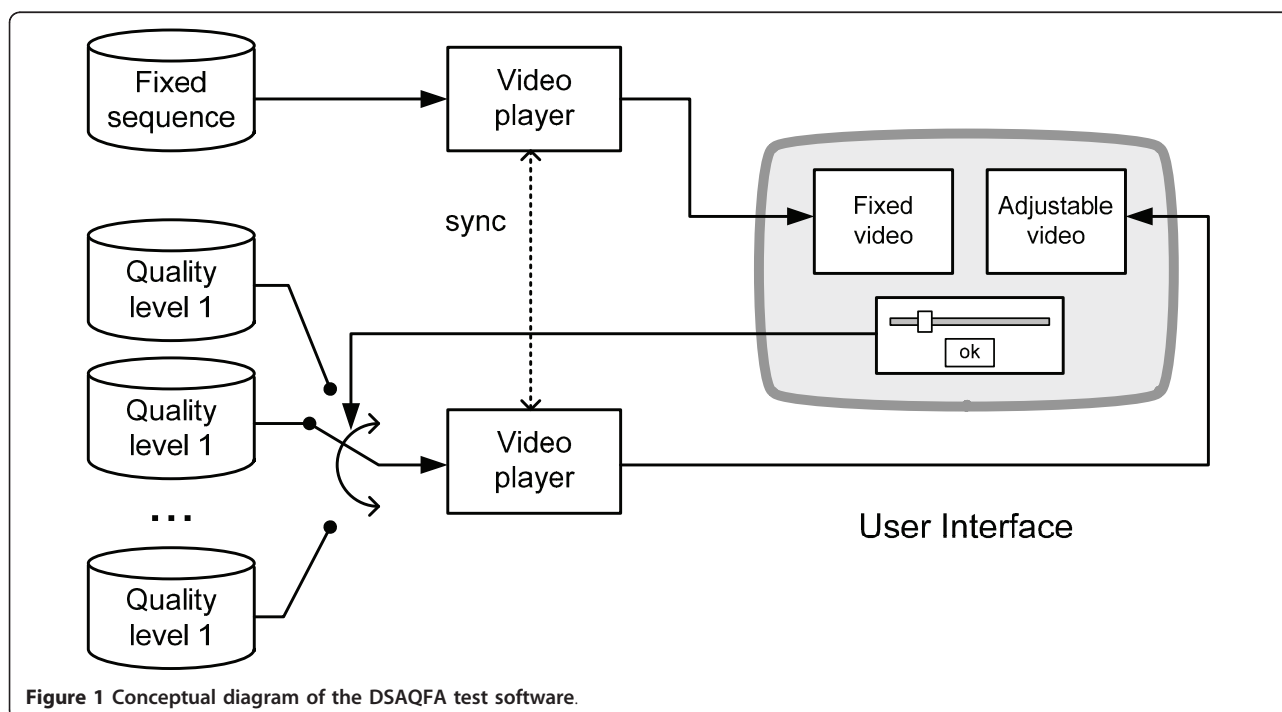

Figure 1 Conceptual diagram of the DSAQFA test software.

sequences are the same. In reality, because we are comparing apples and oranges as described above, the quality of fixed and adjustable video can never be same, hence the quality indices will be close to each other rather than identical. If they are not, it indicates that the objective metric used is not capable of giving reliable relative quality estimates between/across different types of distortions applied to the same content.

3.2. Practical experiments

Two subjective quality assessment studies have been carried out, employing the DSAQFA method. In both studies, the intention was to compare the relative performance of objective quality metrics when sequences may contain either source distortion or both source and channel distortions. Figure 2 illustrates how the distorted sequences were generated. Different levels of source distortion were generated by encoding the original raw video sequence with H.264/AVC, using different quantization parameters (QP). Channel distortion was generated by injecting bit errors to the encoded bitstreams. To simulate the bursty distribution of bit errors in a realistic radio channel, we have used the well-known Gilbert-Elliot model [17]. Since the H.264/AVC standard does not intrinsically support bit error resilience, we have employed a robust packetization scheme in our channel simulation and decoded the erroneous sequences with a modified version of the H.264/AVC reference codec with improved capability to handle bit errors. The details of the packetization scheme and bit error handling are out of the scope of this article, but interested readers may refer to [1,18].

The first assessment study has been reported originally in [7]. In this phase, we had ten different content files, each in CIF resolution (352 × 288 pixels). For each content, there were 4 different fixed impairments and 16 adjustable impairments (see Table 1). The four different

fixed sequences contained different combinations of source distortion (defined by QP) and channel distortion (defined by bit error rate, BER). The 16 different adjustable sequences contained different levels of source distortions (i.e., each one encoded with different QP). This resulted in 40 fixed sequences (10 contents × 4 fixed QP levels) and 160 adjustable sequences (10 contents × 16 fixed QP levels).

In the second study, reported first in [8], the roles of the fixed and adjustable sequences were swapped: the level of channel distortion was adjusted instead of source distortion. In that study, we used four different contents, which had been identified in the first study as being representatives of separate content classes giving distinctively different results. Sample frames of these contents are shown in Figure 3. Six different combinations of QPs were used for the fixed and adjustable sequences (see Table 2). In total, there were 12 fixed sequences (4 contents × 3 fixed QP levels) and 132 adjustable sequences (4 contents × 3 fixed QP levels × 11 BER levels).

4. Results

In this study, our main intention was to evaluate the relative performance of different quality metrics, when different video quality distortion types are concerned. Since compression artifacts with fixed parameters are typically spread evenly along the content both spatially and temporally, we may consider source distortion a more generic type of distortion than channel distortion. Most of the existing objective quality metrics have been evaluated using primarily sequences with source distortion. However, for a metric to be widely applicable in a real-application scenario, it should give the same results for any two sequences that are subjectively perceived as equally disturbing or pleasant, regardless of distortion type. The subjective experiments described in Section
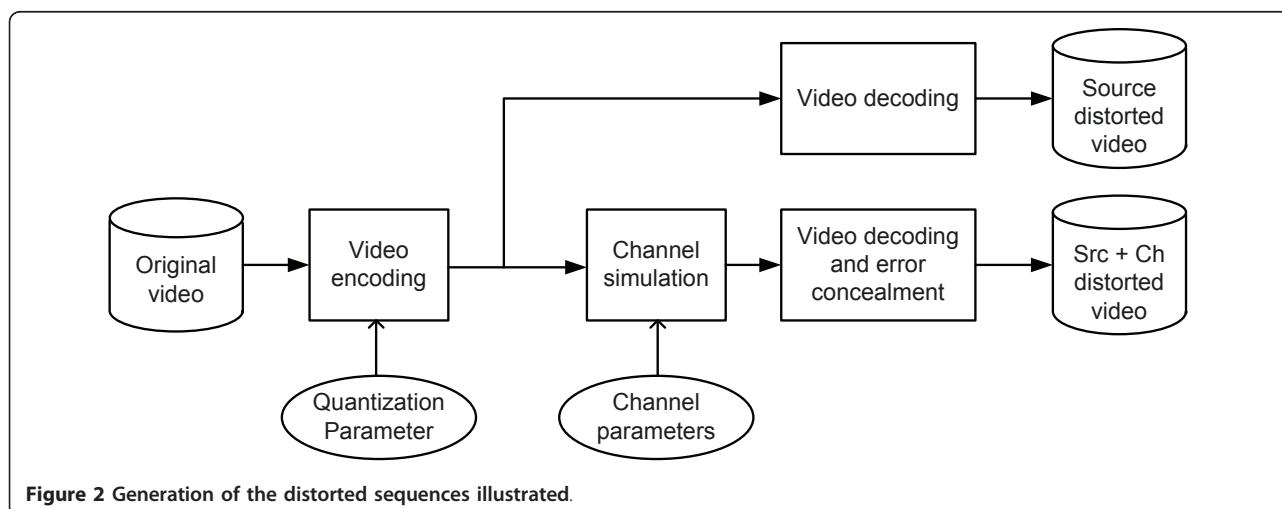

Figure 2 Generation of the distorted sequences illustrated.

Table 1 Test configurations in the study published in QoMEX 2009 [7]

Fixed	QP = 24 BER = 6.5×10^{-5}	QP = 35 BER = 6.5×10^{-5}	QP = 35 BER = 2.7×10^{-5}	QP = 42 BER = 6.5×10^{-5}
Adjustable	QP = 24-51 BER = 0	QP = 24-51 BER = 0	QP = 24-51 BER = 0	QP = 24-51 BER = 0

Ten different content files were included ('Akiyo', 'Bus', 'City', 'Coastguard', 'Football', 'Hallmonitor', 'Harbour', 'Ice', 'Mobile', and 'Soccer').

3.2 allow us to compare the objective quality value of the fixed sequence against the average quality value of the sequences chosen by the test subjects, representing a sequence subjectively perceived as equally good or bad. In [7,8], we have used PSNR values as objective quality values, but in the following, we present the respective results with four additional quality metrics, including two video quality metrics: the general model of VQM proposed byPinson and Wolf [2], MOVIE [19], and two image quality models: structural similarity (SSIM) model [20], and a no-reference image quality metric (NRIQM) [21]. Temporal averaging over all the frames is used with SSIM and NRIQM to measure video quality.

The VQM model extracts several features from the video sequences, expressing spatial gradient activity, chrominance information, contrast information, and absolute temporal information. They are compared using functions modeling the visual masking of the spatial and temporal impairments. Since motion information is the

a) Akiyo b) Bus

c) Harbour d) Ice

Figure 3 Sample frames from the four sequences representing different content classes. **(a)** Akiyo **(b)** Bus **(c)** Harbour and **(d)** Ice.

Table 2 Test configurations in the study published in IMSAA 2009 [8]

Fixed	QP = 48 BER = 0	QP = 48 BER = 0	QP = 48 BER = 0	QP = 42 BER = 0	QP = 42 BER = 0	QP = 38 BER = 0
Adjustable	QP = 42 BER = 0-1.5 × 10^{-2}	QP = 35 BER = 0-1.5 × 10^{-2}	QP = 24 BER = 0-1.5 × 10^{-2}	QP = 35 BER = 0-1.5 × 10^{-2}	QP = 24 BER = 0-1.5 × 10^{-2}	QP = 24 BER = 0-1.5 × 10^{-2}

Four different content files were included ('Akiyo', 'Bus', 'Harbour', and 'Ice').

most important clue in video, it can also be used to emphasize the degradation of spatial and temporal fidelity in the distorted video sequence in comparison to the original sequence.

Seshadrinathan and Bovik [19] proposed to evaluate the motion quality along computed motion trajectories in the MOVIE quality model. Video quality can also be measured by pooling the quality values of individual video frames temporally. Using this approach, two image quality metrics were also included in our experiments. Under the assumption that the human visual system (HVS) is highly adapted to extract structural information from the field of vision, Wang et al. [20] proposed that measuring the change of structural information can provide a good approximation of the perceived image distortion. The SSIM index is used to measure image distortion based on the comparison of luminance, contrast, and structure between the original and distorted images. Furthermore, because the original video sequence is typically not available as a reference in a realistic communication application scenario that is relevant for our study, we tested an NRIQM as well. Because some of the artifacts introduced to the video sequences in our experiments are similar to those artifacts caused by JPEG compression, we have chosen an NRIQM that was proposed for measuring the quality degradation caused by JPEG compression [21].

These quality metrics were performed either between the distorted sequences (fixed or adjustable sequences) and the original, undistorted sequences (for PSNR, VQM, MOVIE, and SSIM), or on the distorted sequences alone (for NRIQM), in order to objectively evaluate the quality of the fixed and adjustable sequences against the original versions. To allow comparison between different metrics with different scales, we have first mapped all the values into a normalized range from 0 to 1, where 0 represents the minimum and 1 the maximum observed value. The mapping from original quality value Q to the normalized value Q' is done using Equation 1, where MIN and MAX are the minimum and maximum raw quality indices observed in our data combined from the experiments in [7,8]. MIN and MAX values for each metric are summarized in Table 3.

$$Q' = \frac{Q - \mathrm{MIN}}{\mathrm{MAX} - \mathrm{MIN}} \qquad (1)$$

In order to evaluate the overall performance of each metric, we have computed the root mean squared error (RMSE) between the normalized quality values of the fixed sequences and the normalized mean quality values of the adjustable sequences in each test case. The numerical results are summarized in Table 4. Surprisingly, the best combined result (i.e., lowest RMSE) is obtained using PSNR as a quality metric. With most of the metrics, the difference between the RMSE results from experiments in [7,8] is relatively large, and more experiments should be conducted to form a more robust evaluation of the metrics. However, as seen in the results, PSNR is among the two best metrics in both subsets of test cases, and it beats all the other metrics, except for SSIM, with a clear margin.

In the next phase, we evaluated the direction of bias in the measured channel distortion. With the basic assumption that source distortion is the generic distortion and channel distortion is a special type of distortion, we can presume that the quality value for the sequence with source distortion Q_{SRC} represents the comparison point of the quality, and the quality value for the sequence with both source and channel distortions $Q_{\mathrm{SRC+CH}}$ is biased due to the insufficient capability of the used metric to model the perceived impact of channel artifacts. Figure 4 shows the average difference diff between Q_{SRC} and $Q_{\mathrm{SRC+CH}}$, computed for four different content types ('akiyo', 'bus', 'harbour', and 'ice'). To allow comparison between metrics, diff has been computed using the normalized quality indices Q'. Equation 2 shows the formula for computing diff, where n is the total number of test cases for the content in question. In case of VQM, MOVIE, and NRIQM, the sign of diff has been switched, since they measure distortion rather than quality (i.e., lower value represents higher quality). Therefore, a positive value of diff indicates that perceptual impact of channel distortion is overestimated, negative value that it is underestimated. The values shown in Figure 4 have been obtained using the data from both experiments [7,8], and

Table 3 Minimum and maximum observed raw quality indices

Metric	PSNR	VQM	MOVIE	SSIM	NRIQM
MIN	25.44	0.0577	0.0201	0.842	22.44
MAX	42.02	0.8218	1.1178	0.996	72.30

Table 4 RMSE between measured quality values of fixed and adjustable sequences

Metric	PSNR	VQM	MOVIE	SSIM	NRIQM
RMSE [6]	0.115	0.160	0.219	0.081	0.187
RMSE [7]	0.081	0.303	0.227	0.164	0.299
RMSE all data	0.103	0.224	0.222	0.119	0.235

the total number of test cases per content was 10 (four from [7] and six from [8]).

$$\text{diff} = \frac{\sum_{i=1}^{n} Q'_{\text{SRC}}(i) - Q'_{\text{SRC}+\text{CH}}(i)}{n} \qquad (2)$$

As shown in Figure 4, the bias of estimating the perceptual impact of channel distortion depends on the metric and the content. PSNR, VQM, MOVIE, and SSIM may either under- or overestimate the quality. Contrastingly, NRIQM seems to systematically underestimate the perceived quality degradation related to channel errors; in other words, the quality values produced by this metric give a too positive impression of the perceived quality, when channel distortion is present.

The results with the additional metrics support the general conclusion based on the results with PSNR [7,8]: the perceptual impact of channel distortion is often either under- or overemphasized by an objective metric, and the direction of the bias depends on the content. 'Akiyo' represents a content type of low spatial and temporal activity, and for this type of content all the tested metrics tend to underestimate the perceptual impact of channel distortion. For static sequences like this, transmission errors attract

the viewer's attention easily, and the objective metrics studied here apparently fail to capture this effect. Contrastingly, 'bus' sequence represents the opposite type of content with high spatial and temporal activity. For this type of content, the bias is reversed with all the metrics studied here, except NRIQM. This is understandable by intuition, since on-screen activity, motion, and multitude of details efficiently mask the impact of transmission errors. 'Harbour' represents content with low motion level but a lot of spatial details, and 'ice' content with high motion but low spatial activity level. For these two contents, the bias is in most cases smaller than for the first two content types.

5. Discussion

Several related studies have been conducted to compare the performance of different objective video quality models [22-24]. Typically, the metrics are evaluated by computing the correlation between the subjective MOS and the objective quality index. Since different scales are employed in different quality metrics, the objective quality indices are first mapped into a scale that is similar to the subjective MOS scale, using regression analysis (usually nonlinear fitting). Even though these studies provide valuable information of the performance of different quality metrics, more research is still needed to construct a more conclusive view to the usability of these metrics in different application scenarios. Subjective MOS itself may form a potential source of uncertainty: it is well known that subjective scores are influenced by several factors not directly related to video quality itself, such as test arrangements, physical environment, viewer expertise, and cultural
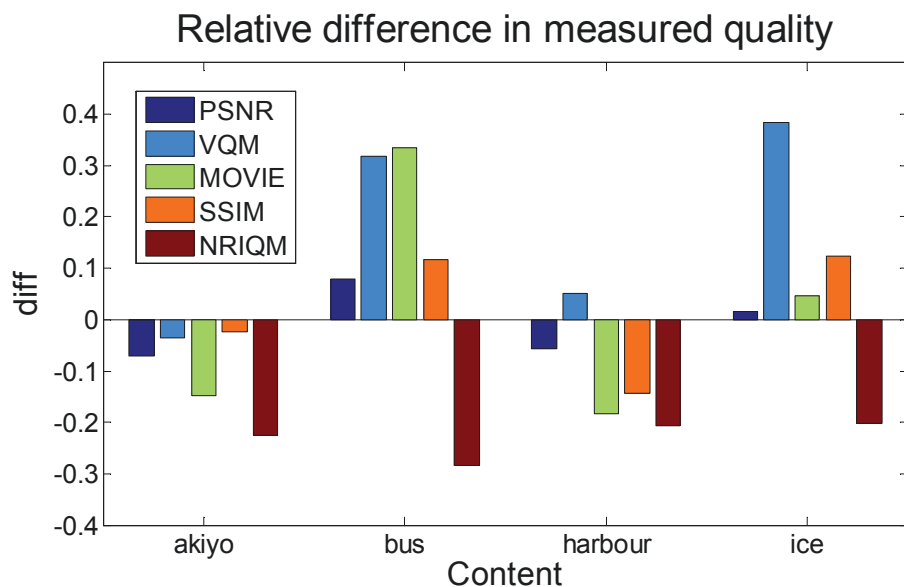

Figure 4 Average difference between objective quality values of perceptually equal sequences.

background. Subjective quality assessment studies conducted at different laboratories have shown to give different absolute MOS results, even if exactly the same testing procedures are followed. This is why a subjective MOS value cannot be considered as absolute indicator of the quality [25]. However, subjective MOS values can still provide useful information about the relative quality of a video sequence with respect to another sequence within the same context.

When the validity of a comparison study is assessed, the application scenario must be carefully considered. Kotevski and Pece [22] have used video sequences with artifacts such as hue, saturation, and Gaussian noise, which are not relevant for a typical real-life multimedia communications scenario. Loke et al. [23] have compared objective metrics subjected to H.264 and MPEG-4 video compression at different bitrates. Therefore, the results are useful for evaluating how well the studied objective metrics assess the impact of source distortion, but the results cannot be generalized for channel distortion. On the other hand, the study by Seshadrinathan et al. [24] includes sequences with both source and channel distortions, but their dataset combines sequences with several different contents. It is well known that PSNR values are highly sensitive to content [4], and bad performance of PSNR in that kind of experiment is therefore expected. In contrast, our study has revealed that in a scenario with fixed content and different distortion types (source versus channel distortion), PSNR is actually capable of comparing the relative quality more accurately than more advanced metrics performing better in the more generic scenarios. This should not be interpreted as evidence of a good performance of PSNR, but rather as evidence that even the more advanced metrics have their weak points. Therefore, more research efforts are still required to develop objective video quality metrics with a good performance across a wide range of applications.

Subjective quality assessment based on paired comparisons can easily become impractical, when combinations of two (or more) different quality factors (e.g., distortion types) are involved in different proportions. We have shown that DSAQFA can substantially facilitate the subjective assessment process in these kinds of scenarios. In the related work, Loke et al. [23] have used impairment scaling with DSIS method variant II and Seshadrinathan et al. [24] have used continuous MOS scaling in a single stimulus test arrangement. These methods are feasible for such studies, but they do not solve the uncertainties related to quality ratings, as discussed in Section 2.2. In addition, DSIS II method leaves little time for the test subjects to give their scores, and in single stimulus method the quality assessment task is demanding due to the lack of the reference. This is why we believe that our

experiences with DSAQFA provide a useful contribution in the field of subjective video quality assessment. The study by Kotevski and Pece [22] is based solely on the numerical analysis of objective measurements, and it is therefore methodologically not comparable with our study.

In this article, we have used DSAQFA to study the relative performance of different objective quality metrics across different distortion types, in this case source and channel distortions. However, it is possible to use these results even further to develop more accurate objective metrics or adjust the existing metrics. The first steps into this direction have been taken in our related study, where we have analyzed numerically the bias in PSNR values toward either channel or source distortion [26]. Since the bias is dependent on the content type, we have tried to estimate the bias from parameters describing the content type, in particular spatial and temporal activity levels. In this way, we have been able to improve the accuracy of predicting the relative perceived quality levels from PSNR in the presence of source and channel distortions [26]. In the future, further improvements could potentially be gained by extending this approach to other objective metrics with better baseline performance than PSNR. In addition, the bias could possibly be predicted more accurately by finding more relevant parameters for content classification.

6. Conclusions
In this article, we have used estimates of mutually respective quality levels for sequences with qualitatively different types of distortions, obtained from a novel subjective quality assessment method that is suitable for comparing different types of quality distortions. The goal of this study was to find out how well different objective quality metrics work when the intention is to evaluate the relative quality levels of video sequences affected by qualitatively different artifacts, namely source and channel distortions.

This study is an extension of our earlier work, in which we have observed that channel errors in sequences with intensive temporal activity seem to have greater impact on PSNR than on the perceived subjective quality. In this article, we have included four other objective metrics in addition to PSNR. The results suggest that the content-dependent tendency of either over- or underestimating the perceptual impact of channel distortion is also present in other metrics, and in most cases, the direction of the bias is the same for all the metrics. An exception is the NRIQM, which systematically underestimates the impact of channel distortion, regardless of content. We assume that our results will be useful for improving the capability of objective metrics to measure the relative perceived distortion originating from different causes, most essentially compression and transmission errors.

Author details
[1]Centre for Quantifiable Quality of Service in Communication Systems (Q2S), Norwegian University of Science and Technology, O.S. Bragstads plass 2E, 7491 Trondheim, Norway [2]Department of Photonics Engineering, Technical University of Denmark (DTU), Oersteds Plads, bldg 343, 2800 Kgs. Lyngby, Denmark

Competing interests
The authors declare that they have no competing interests.

References
1. Korhonen J, Perkis A: **Wireless congestion control based on delivery of erroneous packets.** *Proc. VCIP'09* San Jose, California, USA; 2009.
2. Pinson MH, Wolf S: **A new standardized method for objectively measuring video quality.** *IEEE Trans. Broadcast* 2004, **50(3)**:312-322.
3. Wang Z, Bovik A: **Why is image quality assessment so difficult?** *Proc ICASSP'02* Orlando, Florida, USA; 2002.
4. Hyunh-Thu Q, Ghanbari M: **Scope of validity of PSNR in image/video quality assessment.** *Electron Lett* 2008, **44(13)**:800-801.
5. Mu M: **An interview with video quality experts.** *ACM SIGMM Records* 2009, **1(4)**:1-10.
6. ITU-R: **Methodology for the subjective assessment of the quality of television pictures.** *ITU-R Recommendation BT.500-11* Geneva, Switzerland; 2004.
7. Reiter U, Korhonen J: **Comparing apples and oranges: subjective quality assessment of streamed video with different types of distortion.** *Proc QoMEX'09* San Diego, CA, USA; 2009.
8. Korhonen J, Reiter U: **Analysis on the perceptual impact of bit errors in practical video streaming applications.** *Proc IMSAA'09* Bangalore, India; 2009.
9. Lazarus M: **The great spectrum famine.** *IEEE Spectr* 2010, **47(10)**:26-31.
10. Welzl M: **Passing corrupt data across network layers: an overview of recent developments and issues.** *EURASIP J. Appl. Signal Process* 2005, **2005(2)**:242-247.
11. Larzon L, Degermark M, Pink S: **UDP Lite for real time multimedia applications.** *HP Technical Report HPL-IRI-1999-001* 1999.
12. Singh A, Konrad A, Joseph A: **Performance evaluation of UDP lite for cellular video.** *Proc. NOSSDAV'01* Port Jefferson, New York, USA; 2001, 117-124.
13. Khayam S, Karande S, Radha H, Loguinov D: **Performance analysis and modeling of bit errors and losses over 802.11b LANs for high bit-rate real-time multimedia.** *Signal Process. Image Commun* 2003, **18(7)**:575-595.
14. Masala E, Bottero M, De Martin JC: **Link-level partial checksum for real-time video transmission over 802.11 wireless networks.** *Proc of VTC'05-Spring* Stockholm, Sweden; 2005, 2864-2868.
15. ITU-T Rec. BT.500-11, Methodology for the subjective assessment of the quality of television pictures, International Telecommunication Union, Geneva, Switzerland. 2004.
16. Watson A, Sasse M: **Measuring perceived quality of speech and video in multimedia conferencing applications.** *Proc ACM MM'98* Bristol, UK; 1998.
17. Wang H, Moayeri N: **Finite state Markov channel–a useful model for radio communication channels.** *IEEE Trans Veh Technol* 1995, **44(1)**:163-171.
18. Korhonen J, Frossard P: **Bit-error resilient packetization for streaming H.264/AVC video.** *Proc MV'07* Augsburg, Germany; 2007.
19. Seshadrinathan K, Bovik AC: **Motion tuned spatio-temporal quality assessment of natural videos.** *IEEE Trans Image Process* 2010, **19(2)**:335-350.
20. Wang Z, Bovik AC, Sheikh HR, Simoncelli EP: **Image quality assessment: from error visibility to structural similarity.** *IEEE Trans Image Process* 2004, **13(4)**:600-612.
21. Wang Z, Sheikh HR, Bovik AC: **No-reference perceptual quality assessment of JPEG compressed images.** *Proc ICIP'10* Rochester, NY, USA; 2002.
22. Kotevski Z, Pece M: **Performance comparison of video quality metrics.** *Proc SPIE ICDIP'10* Singapore; 2010.
23. Loke M, Ong P, Lin W, Lu Z, Yao S: **Comparison of video quality metrics on multimedia videos.** *Proc ICIP'06* Atlanta, GA, USA; 2006.
24. Seshadrinathan K, Soundararajan R, Bovik AC, Cormack LK: **Study of subjective and objective quality assessment of video.** *IEEE Trans Image Process* 2010, **19(6)**:1427-1441.
25. Pinson M, Wolf S: **An objective method for combining multiple subjective data sets.** *Proc SPIE VCIP'03* Lugano, Switzerland; 2003.
26. Korhonen J, Reiter U, You J: **On the relationship between perceptual impact of source and channel distortions in video sequences.** *Proc ICIP'10* Hong Kong, China; 2010.