The Performance of Quality Metrics in Assessing Error-Concealed Video Quality

Article *in* IEEE Transactions on Image Processing · March 2020

DOI: 10.1109/TIP.2020.2984356

CITATIONS
5

READS
282

3 authors:

Mohammad Kazemi
University of Isfahan
12 PUBLICATIONS **75** CITATIONS
SEE PROFILE

Mohammed Ghanbari
University of Essex
699 PUBLICATIONS **9,409** CITATIONS
SEE PROFILE

Shervin Shirmohammadi
University of Ottawa
405 PUBLICATIONS **4,668** CITATIONS
SEE PROFILE

The Performance of Quality Metrics in Assessing Error-Concealed Video Quality

Mohammad Kazemi, Mohammad Ghanbari, *Life Fellow, IEEE,* Shervin Shirmohammadi, *Fellow, IEEE*

***Abstract*—In highly-interactive video streaming applications such as video conferencing, tele-presence, or tele-operation, retransmission is typically not used, due to the tight deadline of the application. In such cases, the lost or erroneous data must be concealed. While various error concealment techniques exist, there is no defined rule to compare their perceived quality. In this paper, the performance of 16 existing image and video quality metrics (PSNR, SSIM, VQM, etc.) evaluating error-concealed video quality is studied. The encoded video is subjected to packet loss and the loss is concealed using various error concealment techniques. We show that the subjective quality of the video cannot be necessarily predicted from the visual quality of the error-concealed frame alone. We then apply the metrics to the error-concealed images/videos and evaluate their success in predicting the scores reported by human subjects. The error-concealed videos are judged by image quality metrics applied on the lossy frame, or by video quality metrics applied on the video clip containing that lossy frame; this way, the impact of error propagation is also considered by the objective metrics. The measurement and comparison of the results show that, mostly though not always, measuring the objective quality of the video is a better way to judge the error concealment performance. Moreover, our experiments show that when the objective quality metrics are used for the assessment of the performance of an error concealment technique, they do not behave as they would for general quality assessment. In fact, some newly developed metrics show the correct decision only about 60% of the time, leading to an unacceptable error rate of as much as 40%. Our analysis shows which specific quality metrics are relatively more suitable for error-concealed videos.**

***Index Terms*—Error/Loss concealment, Video quality assessment, Image quality assessment.**

I. Introduction

Due to the tremendous volume of raw video data, compressing the video before streaming it over a network is inevitable. But delivering this compressed video over a best-effort network becomes challenging due to data loss/error during transmission. If the lost/erroneous data cannot be retransmitted, as is the case in video conferencing, tele-presence, or other similar applications, they must be estimated by error concealment techniques. In the past two decades, numerous error concealment techniques have been developed [1]-[3]. To compare the performances of these techniques, one can use objective metrics, which try to evaluate the video quality numerically by signal processing tools. Various objective Image Quality Assessment (IQA) [4]-[5] and Video Quality Assessment (VQA) metrics [5]-[9] are available. Their main goal is to offer a metric which predicts the quality closer to that perceived by the Human Visual System (HVS).

Quality assessment methods can be categorized into three major groups: *full-reference*, *reduced-reference* and *no-reference*. Generally, full-reference methods provide more accurate quality scores; but, when the original source is unavailable, no-reference metrics must be used. To evaluate error-concealed frames and videos, the original sources are available; therefore, we chose full-reference methods for this study.

VQA and IQA metrics are normally used for assessment of various distortion types, such as environmental noise, quantization distortion, blurriness, blockiness, contrast distortion, as well as transmission distortion. Some works have also used them for videos subjected to packet loss. For example, the performances of video quality metrics were examined in the presence of both coding and packet loss artifacts for video communication systems in [10]. A similar work was done in [11] though only using the error concealment technique that comes with the reference software of H.264/AVC. The authors of [12] showed that the Mean Opinion Score (MOS) of the video subjected to packet loss is mostly related to the MOS score of the 10^{th} frame where the quality of the frames show a steady behavior. How to pool the frames' quality to find the video quality was presented in [13]. Quality measurement of the lossy packet video considering its error propagation with a no-reference method was studied in [14], in which the authors estimate the Mean Squared Error (MSE) caused by packet loss. Recently there were some other works for evaluating the quality of corrupted videos, but they focus on frame-squeezing, stalling and rebuffering time [15]-[19]. Detecting impairments in the videos was studied in [60], but the impairments consist of aliasing, combing, compression, false contours/banding, MPEG2/H.264 hits, quantization and upscaling. It should be noted that none of the above work has considered *error concealment* distortion. Therefore, it can be said that the suitability of the metrics for error concealment techniques has not been studied and is challenging because these metrics usually consider quantization distortion, environmental noise, blurriness, etc., while non-ideal error concealment is similar to pixel

M. Kazemi is with the Department of Electrical Engineering, University of Isfahan, Isfahan, Iran (e-mail: m.kazemi@eng.ui.ac.ir).

M. Ghanbari is Professor at the School of Electrical and Computer Engineering, College of Engineering, University of Tehran, Tehran, Iran (e-mail:ghan@ut.ac.ir), as well as Emeritus Professor at the School of Computer Science and Electronic Engineering, University of Essex, Colchester, UK, CO4 3SQ, (e-mail:ghan@essex.ac.uk).

S. Shirmohammadi is with the Distributed and Collaborative Virtual Environments Research Laboratory, School of Electrical Engineering and Computer Science, University of Ottawa, Ottawa, ON K1N 6N5, Canada (e-mail: shervin@eecs.uottawa.ca).

displacement in the frames, and produces a special form of error propagation very different in nature and origin from the former distortion types.

In this paper, an appropriate dataset of error-concealed videos is generated and the capability of 16 existing image/video quality metrics in predicting the human-preferred error concealment technique is studied. In subjective tests, both the standalone error-concealed frame and the whole video clip can be evaluated. The results of our experiments show that the subjective quality of the error-concealed frame alone is not necessarily representative of the video quality containing the lossy frame. For objective tests, while the error-concealed frames scored by IQAs and their corresponding video clips scored by VQAs show almost similar behavior, VQAs' results are mostly and generally closer to reality. This can be explained by the fact that measuring the quality of the error-concealed frames alone does not take into account the role of error propagation. Finally, our results further show that the performance of the existing VQA and IQA metrics in assessing the quality of error-concealed videos is not adequate, and is different from assessing the general quality of the videos distorted by quantization noise, environmental noise, etc. For example, in our results, most of the time Peak Signal to Noise Ratio (PSNR) shows a more reliable performance than some other metrics such as Structural SIMilarity (SSIM).

The rest of the paper is organized in the following order. In Section II, a short review of IQA and VQA algorithms is presented. The methodology of evaluating the metrics is explained in Section III and how to generate the dataset is described in Section IV. The metrics evaluations and analysis of the results are provided in Sections V and VI, respectively. Finally, the paper is concluded in Section VII.

II. Image and Video Quality Metrics

Since the beginning of visual communications, objective quality assessment has been a hot of research, with various IQA and VQA metrics proposed. The quality of images/frames and videos is measured by IQA and VQA metrics, respectively. IQA methods measure only spatial distortion while VQA methods consider also temporal distortion. Even though error-concealed videos have both spatial and temporal distortions, we have used both IQA and VQA methods in our evaluations, because IQA methods are traditionally and popularly used for video where IQA scores of the individual frames of the video are pooled together to find the video's quality score. Our experiments in Section V.B show that some IQA methods are in fact more successful than some VQA methods. Some representative IAQ and VQA methods, used in our work, are described next.

A. IQA methods

The full-reference IQA methods used in this paper are as follows:

PSNR is widely used as a simple metric directly obtained from MSE between the original image and the processed/distorted image. MSE or PSNR is mathematically simple to compute but it does not correlate well with subjective scores [20][21]. However, in [22]-[24] the authors show that as long as the content is not changed, PSNR provides an acceptable judgment. However, PSNR is not useful when different videos with various textures and motion activities are compared.

After PSNR, the next most popular IQA is SSIM [25]. As its name implies, SSIM exploits the fact that HVS is mostly sensitive to the structural information of the scenes. The structural information is gathered and then the first and the second moments of the local parts of the picture are compared in a specific form. SSIM has been extended in various works. In Multi-Scale SSIM (MS-SSIM), the index is obtained considering various resolutions of the images [26]. Feature SIMilarity (FSIM) and its color extension (FSIMc) are the methods which use the structural similarity of low level features [27]. These features are the phase congruency and along with the image gradient magnitude have complementary roles in characterizing the image quality.

Gradient Magnitude Similarity Deviation (GMSD) is proposed in [28], in which the difference between the reference and the distorted images are described in terms of similarity deviation of gradient magnitudes. The pixel-wise gradient magnitude similarity (GMS) between the reference and the distorted images gives a GMS map and pooling these GMS maps using the standard deviation achieves the GMSD quality score.

In [29], using wavelet decomposition, Visual Signal to Noise Ratio (VSNR) is developed by Chandler *et. al*. In another algorithm, the image quality is measured by Separately Evaluating Detail Losses and Additive Impairments (SEDLAI-I) [30]. The detail loss refers to the loss of useful visual information; the additive impairment represents the redundant visual information the appearance of which in the test image will distract the viewer's attention from the useful contents. A metric based on Most Apparent Distortion (MAD) is proposed in [31]. For high-quality images, this algorithm looks for the highly distorted parts, since human subjects do the same. For low-quality images, the distortion is most apparent, and thus the algorithm attempts to look for the image's subject matter.

In [57] a Haar wavelet-based Perceptual Similarity Index (HaarPSI) is introduced. The magnitudes of high-frequency Haar wavelet coefficients are used to define local similarities, and low frequency Haar wavelet coefficients for defining the importance of similarities at specific locations in the image. To capture the horizontal and vertical edges on different frequency scales, six discrete two-dimensional Haar wavelet filters are applied.

In our paper, the above full-reference metrics are used for quality evaluation of error-concealed frames. These image quality metrics can be computed and averaged over all video frames to evaluate the video quality, as discussed in [13]. The IQA scores can be pooled by many methods, such as "Averaging or arithmetic mean", "geometric mean", "Minkowski summation" or many other functions. It is shown in [13] that "Averaging" is better than "Minkowski summation". There exist some other sophisticated approaches

for pooling such as [54][55], but since "Averaging" is very popular, we also use the "Average" IQAs of the frames to find quality score of the video.

It should also be noted that Neural Networks are also used for image quality assessment [40]. But, as surveyed in [59], most of them are (1) no-reference metrics and (2) do not necessarily outperform the classical metrics. As such, we do not consider them in our work.

B. VQA methods

One famous video quality assessment method is Video Quality Model (VQM) [32]. In this method, seven parameters related to the video quality are extracted and are linearly combined to give a VQM quality score.

MOtion-based Video Integrity Evaluation (MOVIE) index is computed based on the structural difference of the spatio-temporal Gabor filtered original and the processed video sequences [33]. A newer version of MOVIE taking into account the flicker, known as Flicker Sensitive MOVIE (FS-MOVIE), is developed in [61].

A simple and effective optical flow-based full-reference VQA algorithm is presented in [34], named FLOw SIMilarity (FLOSIM). The authors use MS-SSIM to measure spatial quality. For temporal quality, some statistical parameters of the optical flow are measured and their deviations from those of the reference video are used for quality metric.

The extension of the method proposed in [30] is also used for video quality assessment in [35][36], which we name (SEDLAI-V). The temporal information, motion-based contrast sensitivity function and visual masking are incorporated to make the metric closer to HVS.

An extension of the MAD algorithm is presented in [37] in which Spatio-Temporal MAD (STMAD) is computed for video quality measurement. In STMAD, the MAD algorithm is applied on spatio-temporal slices of the video and then they are combined appropriately.

The study of lost packet video quality and its error propagation impacts can be found in [38], which proposes to use PSNR Drop Sum (PDS) for video quality assessment. Considering the clipping phenomenon and the forgiveness effect [39], Weighted Modified PDS (WMPDS) is suggested for single loss conditions.

There are also metrics based on fusion of some other metrics. One example is the method presented in [58] where the combinations of PSNR, SSIM, MS-SSIM, FSIM and some other metrics are used as a criterion for block level quality measurement. This algorithm is proposed for parameter selection using rate-distortion optimization in the encoders. A metric known as Video Multi-method Fusion Approach (VMAF) has been introduced by Netflix [63]. Some spatial and temporal features are extracted which are then pooled within each frame to produce one feature value per frame of a video. With Support Vector Machine (SVM) training, these features are aggregated over the video frames.

Full-reference Assessor along Salient Trajectories (FAST) is introduced in [62], in which the motion velocity and motion content along trajectories are extracted and quality degradations on them are computed as the dynamic quality index. For the static index, they have used an IQA to find the spatial quality (e.g. GMSD in [62]). The multiplication of the static and dynamic quality scores is the final score in the FAST metric.

III. Experiment Methodology

To find out which objective method provides the best match with reality, the scores given by the human subjects and those from the objective methods must be compared. An objective method which provides an output closer to the human perceptional quality is more accurate. Therefore, the first requirement is to know the human subjective judgment about the quality.

A. Subjective scores

Based on the guidelines described in [41], there are several methods which could be used for subjectively scoring the videos. Among them Pair Comparison (PC) is appropriate for our purpose. In this method, the test sequences are presented in pairs and the subjects vote which one is better. For example, for three test sequences of A, B, and C, there are six possible combinations: AB, AC, BA, BC, CA, CB, and then a judgment is made to find which element in a pair is preferred by the subjects. Note that A versus B and B versus A are examined separately.

In order to not force the subjects for judgment of difficult comparisons, we also add another vote; namely "No Confidence", used when a subject has difficulty in rank ordering of the clips. Furthermore, as explained in the next sub-section, if there is no dominant vote in the subjects' decisions for a pair of videos, we label the score of that pair as "No Confidence".

B. How to get the subjective scores?

The subjects were undergraduate students in the age range of 20-21 years old, consisting of 12 female and 18 male students. In the field of video processing, they were novice, but they were naturally able to judge which video clip had better perceptual quality. There was no special setup for this study; a lab with normal illumination and a pre-programmed computer showed the paired videos and recorded the subjects' votes. We used a 21.5 inch monitor for our study, and the subjects were free to choose their distances from the screen. Since our evaluation is comparative, the screen size is not a critical issue. The subjects could replay the paired videos several times, but they could not change their votes after the decision.

We ran three sessions for each of CIF, HD and full-HD video clips. So, three subjects could assess their video with their computer without interfering with other sessions. Each session took almost 30 minutes. It does not matter how many clips are voted by the subject during the 30 minutes; the session was transferred to another subject after this time.

The subjects' task was to decide which video from a pair is better perpetually. They did not know which concealment technique has been used for each video clip. For a set of 4

error-concealed videos, there were 12 pairs and hence 12 comparisons were made for each video clip. A software program played the pair of videos in a random order and asked the subjects if the first video is "Better" or "Worse" or "Equal" to the second video, or "No Confidence".

Normally in image/video quality evaluation, the display size, sitting position, lighting condition, etc. affect the quality of experience of the subjects. For example, a video might be "Excellent" in the view of a subject with a smaller screen size, but "Good" with a larger display size. If there are blurriness or incorrectly-adjusted brightness/contrast or similar distortions, or if we wanted to measure the video production quality, the above environmental conditions were important. However, error concealment distortion is naturally different form the above distortions. As can be seen from our Error-Concealed Video Dataset (ECVD) in [50], the concealment distortion is clearly visible in conventionally used monitors and normal environmental conditions. Furthermore, we asked for a comparative judgment form the subjects; i.e., they were required to vote which video is better, and not what is the actual quality of a single video. Our subjects were not asked to vote, for example, which video has "Excellent" quality. Therefore, since the environmental condition is common in each comparison, it does not need to follow the specific setups recommended by standards such as ITU-T P.910 [41] or others.

1. *Subject rejection and "No Confidence" votes*

Our test methodology was designed to make it easy for the subjects to make decision. The subjects were not required to measure the amount of the quality ("Excellent", "Good" and etc.) or even the amount of difference ("Better", "Much Better", "Slightly Better" etc.) They were just asked to select which one is better or if they are equal. The option of "No Confidence" also eases the pressure and makes subjects feel comfortable.

Therefore, there were not many problems commonly faced in the subjective tests such as optimistic or pessimistic behaviors during judgment. One issue we did face was that the subjects may vote too fast and carelessly. We tried to remove such scores with two tests: 1) we included (as recommended by [41]) both AB pair and BA pair in the tests. Each subject votes the quality of video A in comparison to video B for the AB pair, and vice versa for the BA pair. This vote must be consistent for this pair; i.e., if the vote is "Better" for the AB pair, then the vote for the BA pair must be "Worse"; otherwise we remove both votes of this subject for this pair. 2) if A was voted "Better" than "B" and B "Better" than C, then A must be voted "Better" than C, otherwise the subject votes for this clip were removed. Generally, the mathematical *Transitivity* relation between the comparisons must be valid; otherwise the votes of that subject for that video clip were removed. With the above two tests, we removed inconsistent votes for each video clip.

C. Comparison of the objective metrics with the subjective scores

We need to evaluate how much the objective scores correlate with the subjective PC opinions. As described in [43], the rate of classification errors is one parameter that could be used to evaluate the effectiveness of the metrics. A classification error occurs when the subjective and objective metric scores lead to different rankings about a pair of sources.

Based on the subjective and objective rankings, there are nine possible occurrences as listed in Table I. A metric which provides higher Correct Decision Rate (CDR) and lower False Tie and False Differentiation is clearly more matched with reality.

Table I. Nine possible occurrences for subjective and objective pair comparison and types of classification errors

		Based on Subjective scores		
		Better	**Equal**	**Worse**
Based on Objective scores	**Better**	Correct Decision	False Differentiation	False Ranking
	Equal	False Tie	Correct Decision	False Tie
	Worse	False Ranking	False Differentiation	Correct Decision

However, the objective scores do not usually lead to "Equal" quality, since the scores are not exactly the same for two distinct videos, even if they have the same perceived quality. This means that the objective scores might have uncertain accuracy. In other words, a higher objective score does not necessarily lead to a higher perceptual quality. This fact has been observed for a no-reference image quality metric in [56] too.

In order to minimize this uncertainty in the objective scores, the scores are quantized with a given quantization step size (ΔQ). Then, based on the relative values of the quantized scores, "Better", "Worse" or "Equal" classification is carried out for objective scores [43]. It is clear that with varying ΔQ, the objective classification varies too. For example, with very large ΔQ, the scores are quantized to the same value and all the test videos are classified as "Equal" by the objective metric. Therefore, the "False Tie" error increases with larger ΔQ. On the other hand, for small ΔQ, there may be a higher rate of False Differentiation. To select the best value of ΔQ, one suggestion is to set ΔQ such that the Maximum Correct Decision Rate (MCDR) is achieved [44].

D. Difference between our tests and conventional image/video quality assessment tests

The standard methods of subjective testing and relevant standard test scales cannot be used in our work because of the different nature of the problem. This can be explained as follows:

The standard methods usually use 5-level rating quality scales ("Excellent (5)", "Good (4)", "Fair (3)", "Poor (2)" and "Bad (1)") or impairment scales ("Imperceptible (5)", "Perceptible, but not annoying (4)", "Slightly annoying (3)", "Annoying (2)" and "Very annoying (1)") for the subjective tests. For impairment scaling, 7-level comparison ("Much

Better (3)", "Better (2)", "Slightly Better (1)", "The same (0)", "Slightly worse (-1)", "Worse (-2)" and "Much Worse (-3)") can also be used, as in ITU-T P.913 [42], for example. The values in parentheses are the quantitative scores equivalent to these qualitative judgments. They are then averaged over all subjects' scores to obtain MOS or Differential MOS (DMOS). Then, with the commonly used measures; e.g., Spearman Rank Order Correlation Coefficient (SROCC) and Linear Correlation Coefficient (LCC), how much an objective metric is successful in predicting the subjective quality is quantified.

However, in our application we want a comparison of the error concealed video quality in order to find how much the existing I/VQA metrics are successful in determining the better error concealed video, not in how well they correlate to the subjective absolute rating. In this case, using 5-level scaling for the error concealed videos is not meaningful for many videos. For example, it happens in many cases that both videos are "Good", but one is clearly better than the other, though not "Excellent". A 5-level scale would not be able to capture that. Also, the 7-level comparison scale might lead to two problems: (1) it is difficult and less accurate to say that, for example, this video is "Much Better" or "Better" or "Slightly Better" than the other one. Clearly, deciding "Better/Worse" is much simpler and more accurate than deciding how much better. The 7-level scale needs the "how much", but our application doesn't. (2) Lack of matching of the results of the correlation coefficients commonly used in the literature with our application scenario. For example, assume that two videos have a 2dB difference in PSNR; it is possible that in a pair of error concealed video clips, the video with higher PSNR has "Better" quality than the other, while in another pair, it may have "Much Better" quality than the other in the view of subjects. With the measures of SROCC and LCC and other methods computing the correlation coefficient, it is a negative point that 2dB improvement is not always equivalent to "Much Better" quality, for example. Therefore, we have a reduction in the correlation coefficient in this case. However, in the case of comparing the error concealed videos, we will see in Section V that PSNR has truly worked well, since it has predicted the better video correctly, which is also matched with the MCDR criterion employed in our paper.

Therefore, due to the special nature of our problem, our testing and evaluation methodology is more appropriate and accurate than if we used standard methods of subjective testing and the relevant standard test scales.

IV. Dataset preparation

Some lossy video datasets are publicly available, such as: *EPFL/PoliMI Video Quality Assessment Dataset* [45], *IRCCyN/IVC SD RoI Dataset* [46], *IVP Dataset* [47], *LIVE Video Quality Dataset* [48], and *Poly@NYU Packet Loss (PL) Dataset* [49]. However, none contain videos with different error concealment performances, which is required for our work. Therefore, we created the required ECVD dataset [50] ourselves from 18 CIF, HD and full-HD video sequences as base to generate 120 short video clips, each one error-concealed with 4 different algorithms, making a total of 480 distorted video clips. In the next sub-sections, details of this dataset are provided.

A. The loss policy

For an HEVC coded video, a packet contains an integer number of slices, and each slice contains an integer number of Coding Tree Units (CTUs). Therefore, the result of a packet loss can be corruption of several CTUs in the frames.

To generate loss, some CTUs of one frame in these video clips as shown in Fig. 1 were intentionally dropped. The height of each black (lost) band is one CTU, since in actual scenarios the loss is usually limited to one line of CTU. The losses are in the middle of the frame, denoted as Loss Pattern 1 (LP1), or randomly within the frame, denoted as Loss Pattern 2 (LP2). We used LP1 since the middle part usually has higher motion activity and is challenging for error concealment techniques, and usually has noticeable amount of error-concealed distortion. In LP2, as shown in Fig. 1, the lossy area for HD and Full-HD sequences is distributed within the frame. In order to assist the subjects to make decision faster, the location of lost CTU bands are available to the subjects; if they want to see. This way, they can focus to the concealed area and make comparison easier and faster. We investigated both LP1 and LP2 for HD and Full-HD videos. For CIF sequences, we applied only LP1, since the frame is small and almost 50% of the content is covered with LP1; therefore, there is no need to apply and check the results of LP2.

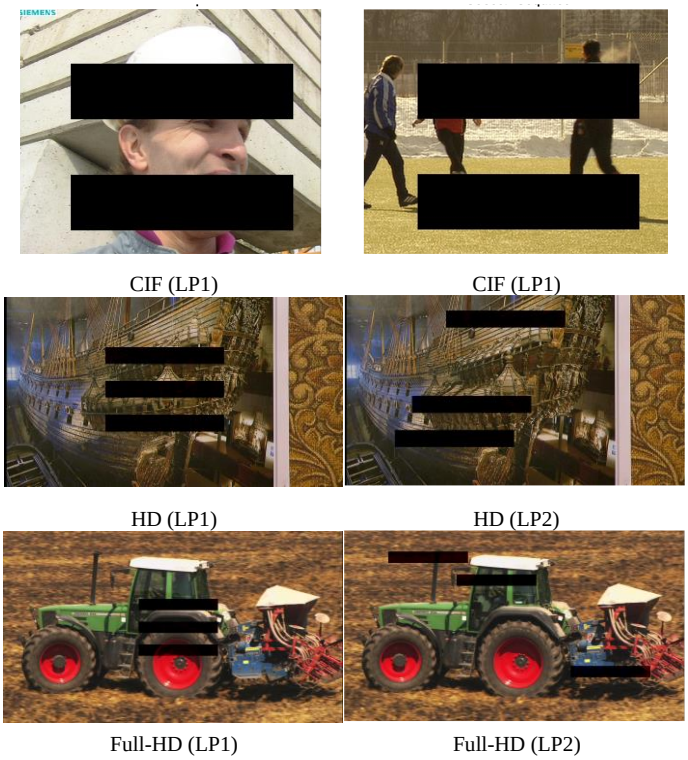


Fig. 1 The applied LP1 and LP2 loss patterns on CIF, HD and Full-HD sequences.

Loss pattern: reality versus experiment

In reality, lossy channels cause packet loss leading to several missed CTUs. In common loss rates, we usually have one or more missed packets in a frame, especially for HD and full-HD sequences. The higher the bitrate/resolution, the more the number of lost packets in a frame at a given packet loss rate. For example, if all frames have the same loss pattern of LP1 or LP2, this is equivalent to 9.3% and 4.1% loss rates for HD and full-HD sequences, respectively. However, for our experiments, applying loss to all frames has two negative impacts: 1) the video might become too distorted for human subjects to compare between the outputs of the concealment techniques; and 2) the quality of the video might switch (Better/Worse) when the video is playing. To demonstrate these facts, the performance curves of 4 error concealment techniques, which are introduced in the next sub-section, are provided in Fig. 2 when LP2 is applied to (a) all frames of a clip of Mobcal HD and Touchdown_pass Full-HD sequences (Fig. 2(a)) and (b) a single frame of those clips (Fig. 2(b)). It can be seen that for the former case we have two problems: (1) the video quality is low especially for Mobcal, and (2) switching the quality; e.g., for Mobcal, AECOD is the best method after the 60th frame, while it was the worst for the earlier frames; or for Touchdown_pass, MVcopy switches multiple times. While these conditions can be easily solved for the objective metrics; e.g. with averaging the quality scores of the frames, human subjects cannot compare the quality of such videos explicitly. In contrast, with one lossy frame only, as reflected in Fig. 2(b), these issues are solved or mitigated significantly. Therefore, in our experiments, we make only one frame lossy in each clip.

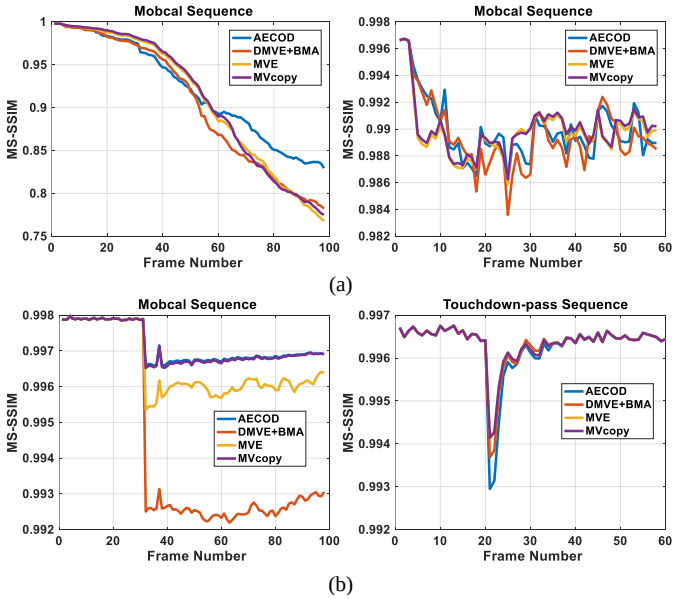


Fig. 2 Comparison of MS-SSIM curves where LP2 is applied (a) to all-frames (b) only to the frame at the 1/3rd point of the clip

To make the subjects' eyes familiar with the content, the lossy frame should not be in the early frames, and in order to have enough frames for error propagation, it should not be in the last frames. Statistically, it does not matter if the lossy frame is selected randomly or it is predetermined, but it should meet the above mentioned conditions. In our experiments, the lossy frame is selected in the midpoint for LP1, and immediately after the 1/3rd point of the video clips for LP2. For CIF sequences, the 10th frame of a clip is the lossy frame. All video clips continue for at least 1 second after the error-concealed frame. The error propagation usually remains no longer than this time, except in rare cases.

Our test video clips are short (about 2 seconds) compared to the video clips in the datasets of [45]-[49]. Since there is one lossy frame in each video clip, the error concealment propagation effect is usually disappeared after a while, due to intra coding. Hence in longer video clips (e.g. conventional 8-10 seconds), either the error concealment quality might be missed or forgotten by the subjects, called forgiveness effect [39], or the good quality interval after the erroneous one (or vice versa) may bias the score due to recency effect [69].

B. Video properties

In error concealment scenarios, we are dealing with two distortions: 1) the distortion in the lossy frame caused by non-ideal error concealment, and 2) the error which propagates to the next frames due to inter prediction. The videos in our dataset have both of the above distortions.

Note that the video clips may originally have some non-idealities. For example we applied a quantization parameter of 22, so the clips have a slight amount of quantization distortion. But this is not an issue because, first, these distortions are easily ignorable as can be visually inspected from our ECVD dataset in [50] and second, distortions other than error concealment are regarded as common mode distortions; i.e., all error-concealed videos contain the same amount of these distortions. Therefore, the subjects as well as the objective metrics can easily decide the better error-concealed videos without being affected by the other distortions. Due to these reasons, we are confident that other sources of quality degradations do not negatively or positively affect our results.

As already mentioned, our video clips were generated from CIF, HD and Full-HD sequences. Snapshots of these sequences as well as a short description of these sequences are provided in Fig. 3 and Table II, respectively. Of the 480 distorted video clips, all were subjected to LP1 loss, while only the 320 HD and Full-HD video clips were subjected to LP2.

C. The examined error concealment techniques

For our study, there is no need to check the state-of-the-art concealment techniques, since the goal of this paper is not proposing an efficient error concealment technique, but instead, how to evaluate the performance of an error concealment operation. Therefore, we have selected the following 4 commonly-used techniques for comparison.

Motion Copy (MVcopy): It is a simple yet very effective method. The lost location is replaced by pixels from the previous frame, using the MV of the co-located block.

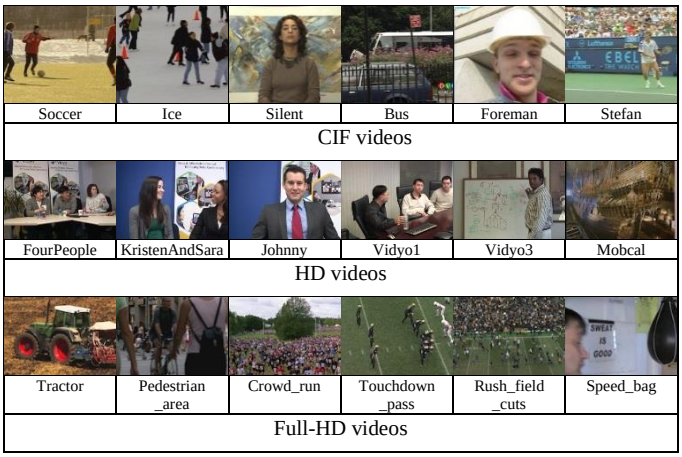


Fig. 3 Snapshots of the sequences used for generation of the test clips

Table II A short description of video sequences used for generation of test clips

Seq. Name		fps	fmt	Content Description
CIF videos (352×288)	Soccer	30	420	Several men playing soccer in a football pitch
	Ice	30	420	Many people play on the iced field
	Silent	30	420	A man moving his hands
	Bus	30	420	A bus quickly moving in the street
	Foreman	30	420	A man in front of a building explains something
	Stefan	30	420	A tennis player in front of spectators
HD videos (1280×720)	FourPeople	60	420	Four people round a conference table
	KristenAndSara	60	420	Two speaking women with static background
	Johnny	60	420	A news reader with static background
	Vidyo1	60	420	Three people in a conference, diagonal view
	Vidyo3	60	420	A standing man writes on a whiteboard
	Mobcal	50	420	Camera pan, toy train moving horizontally with a calendar moving vertically in the background
Full_HD videos (1920×1080)	Tractor	25	420	Camera pan, shows a tractor moving across some fields
	Pedestrian-area	25	420	Still camera, shows some people walking about in a street intersection
	Crowd-run	50	420	Many people running in a park
	Touchdown-pass	30	422	Men playing American football
	Rush-field-cuts	30	422	A rush of the people to the football field
	Speed-bag	30	422	A man hitting a bag, the bag moves very quickly

MV Extrapolation (MVE): In this approach, the MVs of the previous frame block are routed back to the current frame blocks. The MV, the pointed block of which has the maximum overlapping area with the current block, is selected as the recovered MV [51].

Decoder MV Estimation (DMVE) + Boundary Matching Algorithm (BMA): In this method, MVs of the previous frame are not used, and instead the Motion Estimation (ME) is carried out at the decoder. However, since the lost block is not available for ME, the pixels at the boundaries are used for matching [52].

Adaptive Error Concealment Order Determination (AECOD): The basic DMVE is used for MV recovery but the order of concealing the lost 16×16 pixel blocks is determined adaptively [53]. The idea is to first conceal the impact of the lost blocks of the neighborhood blocks with stronger texture intensity, where the texture intensity is measured by standard deviation.

V. Experiments and Evaluations

In this section, the performance of VQA and IQA metrics for comparing the 4 error concealment techniques are evaluated.

A. Subjective quality, error-concealed frame or the video clip?

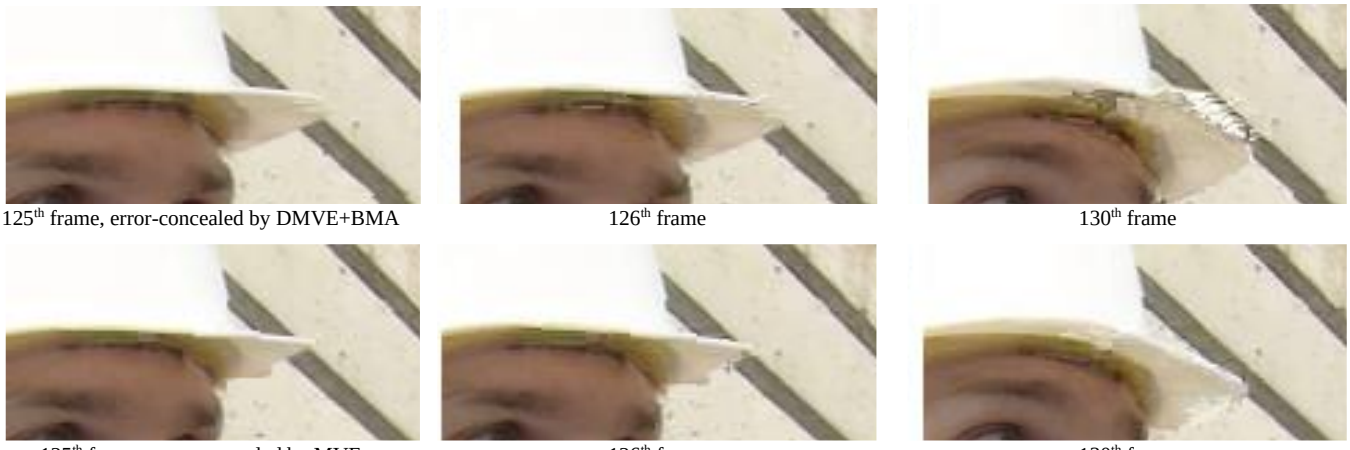
The main goal of a quality metric is accurate prediction of the perceptual or subjective quality. To perform subjective ranking as explained in Section III, we first need to know which source must be judged by the subjects: the error-concealed frame or the whole video clip? Many error concealment research works consider the visual quality of only the error-concealed frame in their evaluation, but is that sufficient? Is it not needed to see the perceptual quality of the video containing not only the error-concealed frame but also the error propagation? To answer this, let us look at some frames from our dataset, as shown in Figure 4.

We can see that looking only at the error-concealed frame (the left most column), DMVE+BMA and AECOD look subjectively better than MVE and MVcopy. However, as we also see from the next frames which are affected by error propagation, the output of the MVE and MVcopy are significantly better, which was not predictable from the error-concealed frame alone. Therefore, it is clear that consideration of the whole video, and not just the error-concealed frame, is needed. This was also the case in [68], which has considered the role of error propagation in its visual quality tests, unlike other existing works, including recent ones [64]-[67]. As such, in all our subjective tests, the whole video clips are subjectively evaluated.

Each lossy frame in each clip was concealed 4 times with 4 error concealment techniques. Each pairs of clips were presented to the subjects. The subjects decided if the output clip of an error concealment technique is “Better”, “Equal” or “Worse”, as given in Table I, or “No Confidence”, as explained before. For each pair, we usually have a dominant vote which is selected as the final subjective ranking. When gathering the votes, there were some rare cases that we had two contradictory votes for a pair, for example “Better” and “Worse”; we labeled this comparison as “No Confidence”. Eventually, we had about 8% “No Confidence” scores for all tests.

B. Computing MCDRs for the metrics

With the source codes provided by the authors, we implemented the metrics explained in Section II. The objective



125th frame, error-concealed by DMVE+BMA | 126th frame | 130th frame

125th frame, error-concealed by MVE | 126th frame | 130th frame

(a) *Foreman* Sequence

331th frame, error-concealed by AECOD | 332th frame | 333th farme

331th frame, error-concealed by MVE | 332th frame | 333th farme

(b) *Rush_field_cuts* Sequence

326th frame, error-concealed by AECOD | 329th frame | 331th frame

326th frame, error-concealed by MVcopy | 329th frame | 331th frame

(c) *Pedestrian_area* Sequence

Fig. 4 Some samples showing that the subjective quality of the solely error-concealed frame is not sufficient, the left column is the error-concealed frames by the captioned technique, the right two columns are the frames which have no packet-loss but they are distorted due to error propagation.

rank ordering is achieved based on the objective scores. As mentioned in Section III.C, "Correct Decision" indicates whether the objective relative score is matched with human comparison. As already discussed, the scores are quantized at ΔQ step and used for rank ordering. The larger steps make the quantized scores more probable to be equal; hence there is more chance to have "Equal" objective quality. The opposite case happens for smaller step sizes.

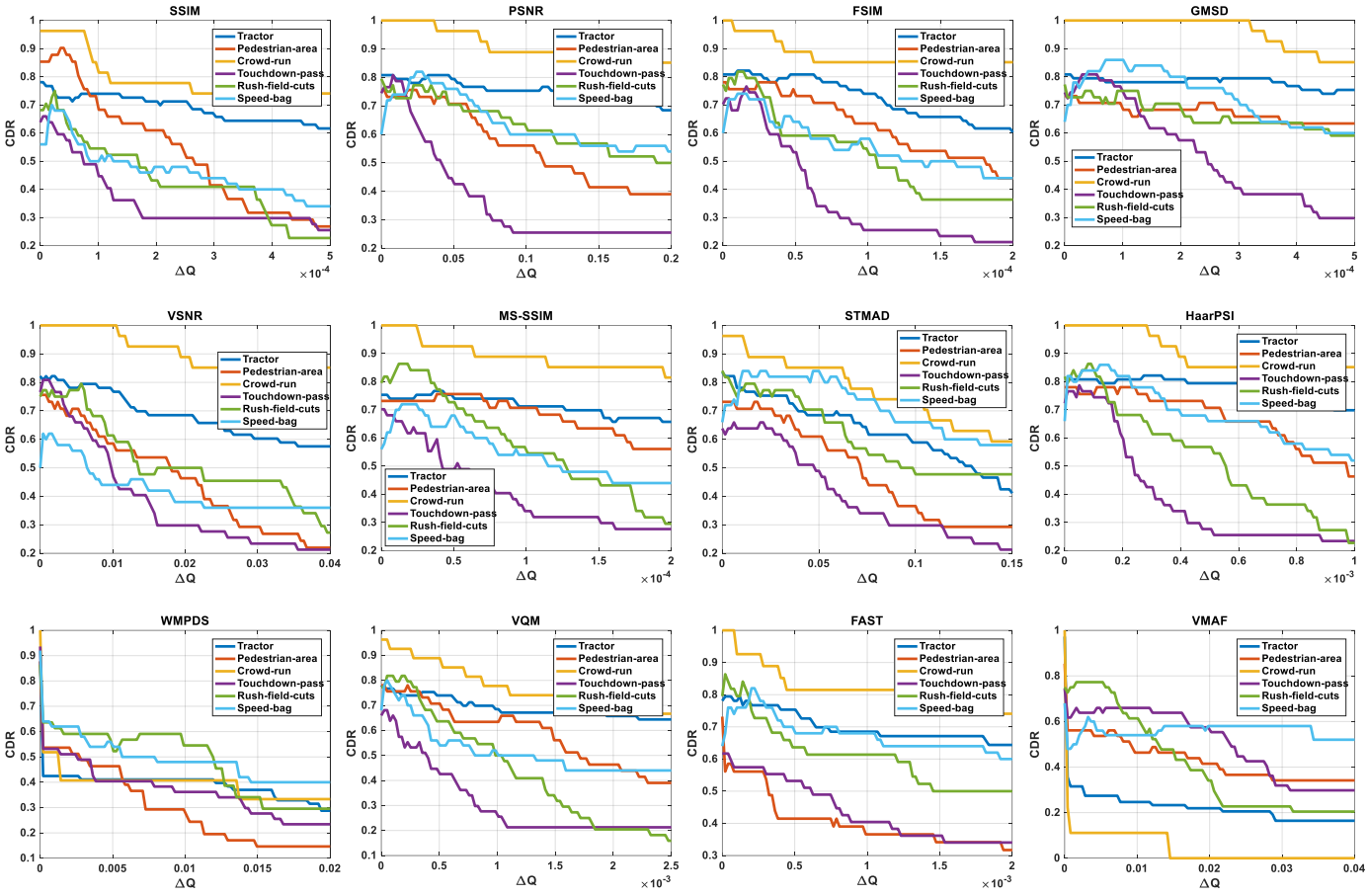


Fig. 5 Vatiation of CDR with ΔQ for varios metrics and full-HD sequence, loss pattern is LP1

Table III *MCDR* based on the subjective scores of the video and objective scores of the error-concealed frames for CIF sequences

	Soccer	***Ice***	***Silent***	***Bus***	***Foreman***	***Stefan***
	(Approach 1)					
SSIM	80.9	88.6	74.5	91.4	68.0	82.9
PSNR	89.4	88.6	74.5	91.4	86.0	77.1
FSIM	89.4	91.4	78.4	91.4	72.0	77.1
FSIMc	89.4	91.4	78.4	91.4	70.0	77.1
GMSD	87.2	88.6	74.5	97.1	82.0	82.9
VSNR	87.2	82.9	70.6	88.6	82.0	74.3
MS-SSIM	85.1	88.6	74.5	97.1	82.0	82.9
SEDLAI-I	80.9	88.6	68.6	88.6	68.0	65.7
MAD	68.1	91.4	70.6	91.4	86.0	82.9
HaarPSI	90.2	90.3	68.9	89.7	84.4	80.0
	(Approach 2)					
SSIM	89.4	88.6	78.4	91.4	76.0	91.4
PSNR	95.7	88.6	76.5	94.3	90.0	82.9
FSIM	95.7	91.4	80.4	94.3	78.0	82.9
FSIMc	95.7	91.4	80.4	94.3	78.0	82.9
GMSD	91.5	88.6	76.5	97.1	88.0	88.6
VSNR	93.6	82.9	72.5	91.4	90.0	77.1
MS-SSIM	91.5	88.6	76.5	97.1	92.0	88.6
SEDLAI-I	87.2	88.6	72.5	91.4	78.0	68.6
MAD	76.6	91.4	72.5	91.4	92.0	88.6
HaarPSI	89.4	91.4	74.5	91.4	90.0	82.9

Fig. 5 shows the variation of CDR with ΔQ when the error-concealed Full-HD video clips are under study. It can be seen that CDR for some sequences such as *Crowd-run* is decreasing. This means that, as we try to make equal the objective scores by the larger ΔQs, the accuracy of the metric in correct comparison is diminished. The reason is that the video clip pairs are not voted as "Equal" for this sequence and with larger quantization step sizes, the correct objective decisions ("Better" or "Worse") are wrongly converted to "Equal". For VMAF, the sensitivity to ΔQ for this sequence is significant; that is, the objective scores are very close to each other, while the other sequences do not have such sensitivity. In other words, a small variation in VMAF leads to noticeable quality variation in *Crowd_run*, while it is not the case for the other metrics and/or sequences. Therefore, we can conclude that the small variation in a metric might or might not be meaningful; it is sequence dependent as this figure shows. From our subjective study, it cannot be determined whether this sensitivity can be assumed as a strength or a weakness of the metric. It needs quantitative rating. However, since the other 12 metrics do not have such sensitivity, the sensitivity of VMAF here is under question, even though its CDR is the best for many small ΔQs.

For some metrics and some sequences, we get higher CDRs with higher ΔQ; maximizing CDR over ΔQ results in MCDR.

Table IV *MCDR* based on the subjective scores of the video and objective scores of the error-concealed frames for HD sequences

	FourPeople		***KristenAnd Sara***		***Johnny***		***Vidyo1***		***Vidyo3***		***Mobcal***	
	LP1	LP2	LP1	LP2	LP1	LP2	LP1	LP2	LP1	LP2	LP1	LP2
SSIM	96.7	93.1	90.0	93.3	96.6	96.4	100	93.3	96.7	96.7	77.8	89.7
PSNR	96.7	93.1	100	96.7	93.1	96.4	100	100	96.7	96.7	92.6	93.1
FSIM	96.7	93.1	96.7	96.7	93.1	100	100	100	96.7	100	96.3	93.1
FSIMc	96.7	93.1	96.7	96.7	93.1	100	100	100	96.7	100	96.3	93.1
GMSD	96.7	96.6	100	96.7	89.7	100	100	96.7	96.7	96.7	92.6	93.1
VSNR	93.3	65.5	76.7	73.3	82.8	75.0	79.2	73.3	90.0	83.3	92.6	86.2
MS-SSIM	96.7	93.1	96.7	96.7	93.1	96.4	100	100	96.7	96.7	88.9	93.1
SEDLAI-I	96.7	86.2	90.0	96.7	96.6	100	100	96.7	96.7	100	92.6	86.2
MAD	96.7	93.1	96.7	93.3	93.1	92.9	95.8	96.7	96.7	93.3	92.6	96.6
HaarPSI	96.7	100	96.7	100	89.7	100	100	100	96.7	100	92.6	100
(Approach 2)												
SSIM	96.7	93.1	96.7	93.3	96.6	100	100	96.7	100	96.7	88.9	89.7
PSNR	96.7	96.6	100	96.7	96.6	100	100	100	100	96.7	92.6	93.1
FSIM	96.7	96.6	100	96.7	96.6	100	100	100	100	100	100	93.1
FSIMc	96.7	96.6	100	96.7	96.6	100	100	100	100	100	100	93.1
GMSD	96.7	100	100	96.7	93.1	100	100	100	100	96.7	92.6	93.1
VSNR	93.3	69.0	76.7	76.7	86.2	75.0	87.5	73.3	93.3	83.3	92.6	86.2
MS-SSIM	96.7	96.6	100	96.7	96.6	100	100	100	100	96.7	88.9	93.1
SEDLAI-I	96.7	89.7	93.3	96.7	100	100	100	96.7	96.7	100	92.6	89.7
MAD	96.7	96.6	96.7	96.7	96.6	92.9	95.8	100	96.7	93.3	92.6	96.6
HaarPSI	96.7	100	96.7	100	93.1	100	100	100	100	100	96.3	100

Table V *MCDR* based on the subjective scores of the video and objective scores of the error-concealed frames for Full-HD sequences

	Tractor		***Pedestrian_ area***		***Crowd_run***		***Touchdown_pass***		***Rush_field_cuts***		***Speed_ Bag***	
	LP1	LP2	LP1	LP2	LP1	LP2	LP1	LP2	LP1	LP2	LP1	LP2
SSIM	75.3	84.3	82.9	91.9	92.6	73.3	68.1	77.6	81.8	84.8	70.0	76.7
PSNR	79.5	88.6	70.7	89.2	100	93.3	68.1	75.5	75.0	63.0	84.0	48.8
FSIM	82.2	85.7	80.5	89.2	100	93.3	72.3	85.7	84.1	47.8	88.0	62.8
FSIMc	82.2	85.7	80.5	89.2	100	93.3	72.3	85.7	86.4	47.8	88.0	62.8
GMSD	79.5	84.3	70.7	91.9	100	86.7	61.7	79.6	84.1	60.9	88.0	60.5
VSNR	82.2	87.1	73.2	81.1	100	90.0	70.2	75.5	75.0	56.5	76.0	37.2
MS-SSIM	78.1	85.7	61.0	86.5	100	96.7	78.7	81.6	84.1	52.2	76.0	53.5
SEDLAI-I	80.8	81.4	48.8	59.5	100	90.0	66.0	65.3	75.0	63.0	84.0	46.5
MAD	82.2	85.7	63.4	81.1	96.3	93.3	72.3	79.6	77.3	60.9	82.0	51.2
HaarPSI	82.2	85.7	82.9	91.9	100	93.3	66.0	81.6	79.5	60.9	86.0	55.8
(Approach 2)												
SSIM	79.5	92.9	85.4	94.6	92.6	80.0	78.7	81.6	90.9	89.1	80.0	81.4
PSNR	82.2	95.7	70.7	91.9	100	96.7	80.9	77.6	84.1	69.6	88.0	53.5
FSIM	84.9	95.7	82.9	91.9	100	96.7	80.9	89.8	93.2	52.2	92.0	69.8
FSIMc	84.9	95.7	82.9	91.9	100	96.7	80.9	89.8	93.2	52.2	92.0	69.8
GMSD	83.6	92.9	75.6	94.6	100	96.7	70.2	83.7	93.2	71.7	92.0	65.1
VSNR	83.6	92.9	75.6	83.8	100	93.3	78.7	81.6	84.1	65.2	78.0	44.2
MS-SSIM	79.5	95.7	65.9	89.2	100	96.7	91.5	83.7	88.6	58.7	84.0	60.5
SEDLAI-I	83.6	88.6	53.7	59.5	100	90.0	76.6	67.3	79.5	69.6	86.0	53.5
MAD	86.3	94.3	65.9	83.8	96.3	93.3	85.1	83.7	86.4	63.0	92.0	58.1
HaarPSI	82.2	92.9	82.9	94.6	100	96.7	72.3	83.7	86.4	69.6	92.0	62.8

The ΔQ at which this MCDR occurs, ΔQ_{opt}, gives the uncertainty of this metric for this sequence, as already mentioned. That is, the subjects have voted "Equal" quality while the scores are different as much as ΔQ.

The mathematical equations for *MCDR* and ΔQ_{opt} are as follows:

$$MCDR = \max_{\Delta Q_{min} \le \Delta Q \le \Delta Q_{max}} \{CDR\ (\Delta Q)\} \quad (1)$$

$$\Delta Q_{opt} = \arg\left\{\max_{\Delta Q_{min} \le \Delta Q \le \Delta Q_{max}} \{CDR\ (\Delta Q)\}\right\} \quad (2)$$

To avoid mapping of the objective scores on each other, ΔQ must be selected sufficiently small; ΔQ_{min} is the maximum ΔQ that has this property. On the other hand, if ΔQ is sufficiently large, all scores are quantized to the same value. The minimum ΔQ that has such property is denoted as ΔQ_{max}.

We used two approaches to obtain the *MCDR*:

Approach 1: A common ΔQ is applied for all clips of a sequence. Each video clip has been error-concealed with 4 techniques, leading to 4 outputs and 6 pairwise comparisons (the subjects do 12 comparisons since we examined both AB and BA pairs) for the video clip. For the *Tractor* sequence, for example, there are thirteen 2-sec videos; therefore, the total number of comparisons becomes 78. The *MCDR* is the maximum CDR obtained when these 78 comparisons are done where a common ΔQ is applied to the scores of all error-concealed clips of the sequence.

Approach 2: As the second approach, we can find ΔQ_{opt} for each video clip. In this approach, associated with each video clip, we have one ΔQ_{opt}; i.e., the optimization of (2) is performed clip-wise and for 6 comparisons. For example, for one clip of *Tractor*, ΔQ_{opt} might be 0.3 dB on PSNR metric, or it might be 0.5 dB for another clip of that sequence. In other words, in this approach, 0.3 dB PSNR might be ignored for one clip of a sequence while it might be important for another clip of that sequence. This approach assists to remove content dependency of the metrics and hence gives higher *MCDR*s. Note that content dependency is not a strength of the metrics, and hence Approach 1 is more strict for metrics evaluation.

Table VI *MCDR* based on the subjective and objective scores of the video for CIF sequences (Approach 1)

	Soccer	***Ice***	***Silent***	***Bus***	***Foreman***	***Stefan***
SSIM	82.9	93.5	71.1	96.6	64.4	86.7
PSNR	90.2	93.5	75.6	93.1	82.2	80.0
FSIM	87.8	93.5	73.3	96.6	66.7	80.0
FSIMc	87.8	93.5	73.3	96.6	66.7	80.0
GMSD	90.2	93.5	73.3	96.6	80.0	86.7
VSNR	87.8	93.5	73.3	93.1	82.2	73.3
MS-SSIM	90.2	93.5	71.1	96.6	73.3	90.0
SEDLAI-V	87.8	93.5	71.1	93.1	77.8	86.7
STMAD	85.4	93.5	75.6	93.1	80.0	90.0
HaarPSI	92.7	93.5	80.0	96.6	82.2	76.7
WMPDS	90.2	93.5	80.0	93.1	82.2	76.7
VQM	90.2	93.5	71.1	93.1	75.6	86.7
MOVIE	87.8	93.5	62.2	93.1	80.0	83.3
FLOSIM	75.6	83.9	46.7	79.3	62.2	63.3
VMAF	87.8	93.5	75.6	96.6	64.4	73.3
FAST	56.1	87.1	55.6	93.1	73.3	86.7
(Approach 2)						
SSIM	92.7	93.5	73.3	96.6	75.6	93.3
PSNR	92.7	93.5	80.0	96.6	91.1	86.7
FSIM	92.7	93.5	75.6	96.6	71.1	86.7
FSIMc	92.7	93.5	75.6	96.6	71.1	86.7
GMSD	95.1	93.5	75.6	96.6	86.7	93.3
VSNR	90.2	93.5	80.0	96.6	88.9	83.3
MS-SSIM	95.1	93.5	73.3	96.6	86.7	93.3
SEDLAI-V	95.1	93.5	75.6	96.6	86.7	96.7
STMAD	92.7	93.5	80.0	96.6	88.9	96.7
HaarPSI	97.6	93.5	84.4	96.6	91.1	76.7
WMPDS	100	93.5	82.2	96.6	91.1	83.3
VQM	95.1	93.5	71.1	96.6	82.2	93.3
MOVIE	92.7	93.5	68.9	96.6	91.1	90.0
FLOSIM	78.0	83.9	51.1	79.3	73.3	73.3
VMAF	92.7	93.5	80.0	96.6	73.3	76.7
FAST	61.0	87.1	60.0	96.6	82.2	86.7

1. *Judgment based on the objective scores of the error-concealed frame*

In this section, the performance of the IQA methods in predicting which error concealment technique gives higher subjective quality is studied. Having computed the objective scores, the *MCDR*s achieved with these scores and the subjective scores obtained in sub-section A are computed with the above stated Approach 1 and Approach 2. The results are given in Tables III-V. In all tables presenting the results of the HD and Full-HD sequences, the values associated with LP1 and LP2 loss patterns are included, while the tables of CIF have only the values of LP1.

The first thing these tables show is the higher MCDRs achieved through Approach 2, as expected, since in Approach 2 the content dependency of the scores are removed, leading to maximum performance. However, it can be seen that the MCDRs associated with CIF and Full-HD sequences are sometimes disappointing. For example, for SSIM and Speed-bag sequence in Table V, MCDR is about 80% (Approach 2 and both loss patterns); this means that for 20% of the comparisons, the decisions of the subjects are not matched to the decisions of the SSIM metric. PSNR metric performs better for this sequence for LP1 compared to the LP2; the reason is content dependency of the error concealment techniques' performance.

For Vidyo1 and Vidyo3 sequences in Table IV, it can be seen that some metrics such as SSIM, PSNR, FSIM, FSIMc, GMSD and MS-SSIM in Approach 2 give 100% correct decision for LP1. The MCDRs of PSNR and SSIM behave differently for Stefan and Foreman sequences in Table III. There are many published error concealment techniques which claim quality improvement by PSNR measurement, but we can see that for error-concealed frames PSNR does not match perceptual quality for about 30% of the comparisons for Pedestrian_area with LP1, and about 50% for Speed_bag with LP2.

We can accumulate the scores of not only the error-concealed frames, but also the successive error propagation infected frames and measure the performance of the metrics, as described next.

2. *Judgment based on video clip containing the error-concealed frame*

In this sub-section, the average IQA scores of the lossy frame and the 1-sec following frames are used for objective decision. Furthermore, VQA metrics are also evaluated, although the MOVIE and FLOSIM metrics were used for CIF sequences only, because they are very memory demanding, time consuming, and complex. Again, the *MCDR*s with the two approaches are computed and given in Tables VI-VIII.

The results of Tables VI-VIII show almost the same behavior as in Tables III-V, but with higher MCDRs for CIF and Full-HD sequences. In the following section, the results of the above experiments and measurements are analyzed.

Table VII *MCDR* based on the subjective and objective scores of the video for HD sequences (Approach 1)

	FourPeople		***KristenAndSara***		***Johnny***		***Vidyo1***		***Vidyo3***		***Mobcal***	
	LP1	LP2	LP1	LP2	LP1	LP2	LP1	LP2	LP1	LP2	LP1	LP2
SSIM	96.7	93.1	96.7	93.3	93.1	100	100	93.3	100	96.7	88.9	89.7
PSNR	96.7	93.1	100	100	93.1	100	100	100	96.7	100	96.3	93.1
FSIM	96.7	96.6	96.7	100	93.1	100	100	100	100	100	92.6	89.7
FSIMc	96.7	96.6	96.7	100	93.1	100	100	100	100	100	92.6	89.7
GMSD	96.7	96.6	100	100	89.7	100	100	100	96.7	100	88.9	93.1
VSNR	86.7	37.9	66.7	76.7	75.9	71.4	70.8	56.7	90.0	80.0	88.9	89.7
MS-SSIM	96.7	96.6	96.7	100	93.1	100	100	100	100	100	92.6	93.1
SEDLAI-V	96.7	93.1	100	100	93.1	96.4	100	100	100	100	92.6	93.1
STMAD	96.7	96.6	93.3	100	96.6	100	100	100	96.7	96.7	88.9	93.1
HaarPSI	96.7	96.6	93.3	100	89.7	100	100	100	100	100	92.6	93.1
WMPDS	96.7	100	100	100	100	100	100	100	100	100	96.3	93.1
VQM	96.7	93.1	96.7	100	93.1	100	100	96.7	100	100	85.2	82.8
VMAF	96.7	96.6	100	100	96.6	100	100	100	100	100	88.9	93.1
FAST	96.7	96.6	93.3	96.7	96.6	100	100	96.7	100	100	92.6	89.7
(Approach 2)												
SSIM	96.7	96.6	100	96.7	96.6	100	100	96.7	100	96.7	88.9	89.7
PSNR	96.7	96.6	100	100	96.6	100	100	100	100	100	96.3	93.1
FSIM	96.7	100	100	100	96.6	100	100	100	100	100	96.3	89.7
FSIMc	96.7	100	100	100	96.6	100	100	100	100	100	96.3	89.7
GMSD	96.7	100	100	100	93.1	100	100	100	100	100	88.9	93.1
VSNR	86.7	37.9	66.7	76.7	79.3	75.0	70.8	56.7	90.0	80.0	88.9	89.7
MS-SSIM	96.7	100	100	100	96.6	100	100	100	100	100	92.6	93.1
SEDLAI-V	96.7	96.6	100	100	96.6	96.4	100	100	100	100	92.6	93.1
STMAD	96.7	100	93.3	100	96.6	100	100	100	96.7	100	88.9	93.1
HaarPSI	96.7	100	96.7	100	93.1	100	100	100	100	100	92.6	93.1
WMPDS	96.7	100	100	100	100	100	100	100	100	100	96.3	93.1
VQM	96.7	96.6	96.7	100	96.6	100	100	100	100	100	88.9	82.8
VMAF	96.7	96.6	100	100	100	100	100	100	100	100	88.9	93.1
FAST	96.7	100	100	96.7	96.6	100	100	100	100	100	92.6	89.7

Table VIII *MCDR* based on the subjective and objective scores of the video for Full-HD sequences (Approach 1)

	Tractor		***Pedestrian_Area***		***Crowd_run***		***Touchdown_pass***		***Rush_field_cuts***		***Speed_bag***	
	LP1	LP2	LP1	LP2	LP1	LP2	LP1	LP2	LP1	LP2	LP1	LP2
SSIM	78.1	82.9	90.2	91.9	96.3	73.3	68.1	77.6	75.0	84.8	70.0	74.4
PSNR	80.8	87.1	75.6	91.9	100	80.0	80.9	77.6	79.5	87.0	82.0	65.1
FSIM	82.2	87.1	78.0	91.9	100	83.3	76.6	79.6	84.1	78.3	76.0	72.1
FSIMc	82.2	87.1	78.0	91.9	100	83.3	76.6	79.6	84.1	78.3	76.0	72.1
GMSD	80.8	84.3	73.2	91.9	100	80.0	80.9	81.6	77.3	87.0	86.0	69.8
VSNR	82.2	87.1	75.6	91.9	100	86.7	80.9	77.6	79.5	71.7	66.0	44.2
MS-SSIM	76.7	85.7	75.6	94.6	100	76.7	70.2	89.8	86.4	78.3	72.0	69.8
SEDLAI-V*	83.6	87.1	82.9	89.2	100	83.3	0.0	0.0	0.0	0.0	0.0	0.0
STMAD	83.6	82.9	73.2	89.2	96.3	83.3	66.0	77.6	84.1	78.3	84.0	58.1
HaarPSI	82.2	84.3	78.0	89.2	100	83.3	78.7	73.5	88.6	80.4	88.0	67.4
WMPDS	87.7	91.4	87.8	97.3	100	96.7	93.6	100	86.4	89.1	92.0	83.7
VQM	78.1	88.6	78.0	89.2	96.3	90.0	68.1	73.5	81.8	73.9	80.0	55.8
VMAF	97.3	98.6	85.4	91.9	100	100	74.5	81.6	77.3	84.8	68.0	83.7
FAST	79.5	88.6	73.2	81.1	100	83.3	61.7	85.7	86.4	80.4	82.0	58.1
(Approach 2)												
SSIM	82.2	92.9	90.2	94.6	96.3	73.3	76.6	83.7	79.5	91.3	78.0	79.1
PSNR	84.9	94.3	80.5	94.6	100	86.7	87.2	77.6	86.4	91.3	88.0	72.1
FSIM	84.9	94.3	80.5	94.6	100	86.7	83.0	83.7	88.6	84.8	84.0	76.7
FSIMc	84.9	94.3	80.5	94.6	100	86.7	83.0	83.7	88.6	84.8	84.0	76.7
GMSD	84.9	90.0	75.6	91.9	100	83.3	87.2	89.8	84.1	91.3	92.0	76.7
VSNR	86.3	92.9	80.5	91.9	100	90.0	89.4	81.6	86.4	76.1	70.0	51.2
MS-SSIM	79.5	94.3	78.0	94.6	100	83.3	83.0	93.9	90.9	87.0	80.0	76.7
SEDLAI-V*	87.7	94.3	82.9	89.2	100	90.0	0.0	0.0	0.0	0.0	0.0	0.0
STMAD	86.3	91.4	78.0	91.9	96.3	90.0	76.6	81.6	90.9	80.4	94.0	62.8
HaarPSI	84.9	92.9	82.9	89.2	100	90.0	85.1	77.6	93.2	84.8	94.0	72.1
WMPDS	87.7	95.7	87.8	97.3	100	96.7	93.6	100	90.9	95.7	92.0	86.0
VQM	82.2	94.3	80.5	89.2	96.3	96.7	78.7	75.5	88.6	78.3	90.0	58.1
VMAF	97.3	98.6	85.4	91.9	100	100	87.2	85.7	86.4	91.3	92.0	88.4
FAST	80.8	94.3	75.6	81.1	100	90.0	61.7	89.8	90.9	87.0	88.0	58.1

* the source code of this metric is only for 420 chroma format

VI. Results

In this section, we have provided 4 results by analyzing the above experiments and tables.

1. *The necessity of considering the video for subjective evaluation*

As verified by the samples shown in Fig. 4, the subjective quality of the error-concealed frame is not representative of the subjective quality of the video clip containing that frame. These samples show that the visual quality of the error-concealed frame is not sufficient to judge the performance of the used error concealment technique. Considering that the goal of video error concealment is to improve the perceptual quality of videos and not images, the subjective tests must be conducted for the evaluation of the video clips.

2. *Approach 2 leads to higher MCDR than Approach 1*

As evident from Tables III-VIII, the *MCDR*s in Approach 2 are higher than those in Approach 1. In Approach 2, the optimization of ΔQ is performed clip-wise, while in Approach 1 it is sequence-wise and over all clips of the sequence. The value of ΔQ_{opt} indicates how much difference in scores leads to observable quality differences. The higher *MCDR*s in Approach 2 imply that ΔQ_{opt} is content dependent. This in fact shows that, for many metrics, the amount of variation in objective scores which leads to perceptual quality difference is content dependent, even though this content dependency is rather small for WMPDS.

3. *Considering whole video usually leads to higher MCDRs than the error-concealed frame only*

For the error-concealed frame, we used IQA metrics. For the video clip, in addition to VQA, we averaged the IQA scores computed for all frames of the clip. To quantify this difference, we can define and compute $MCDR_{gain}$ as given by (3).

$$MCDR_{gain} = MCDR_{video} - MCDR_{image} \quad (3)$$

where $MCDR_{image}$ and $MCDR_{video}$ are the values given in Tables III-V and Tables VI-VIII, respectively. The values of $MCDR_{gain}$, associated with the metrics and test sequences, are given in Table IX. Note that $MCDR_{gain}$ is computed for the same metrics or the same family of metrics (e.g. SEDLAI-I and SEDLAI-V or MAD and STMAD, indicated by SEDLAI-V(I) and STMAD(MAD) in Table IX).

The results of this table are sometimes negative, for example for *Crowd_run* and all metrics with LP2 loss pattern, or VSNR and all HD sequences, and a few other metrics/sequences. But most of the time, they are positive, and larger gains are in favor of $MCDR_{video}$. This means that the video clips evaluated by the metrics are closer to the subjective scores, compared to error-concealed frames only evaluated by IQA methods. For example, as Table IX shows, $MCDR_{gain}$ results of SEDLAI-V when checking the *Stefan* and *Pedestrian_area* sequences are significantly improved. For *Rush_field_cuts*, *Speed_bag* with LP2 we also have positive and significant $MCDR_{gain}$ even though it is not the case with LP1. Generally, taking into account the video clips for objective quality measurement leads to better prediction on average, even though the gain is not significant for some cases.

4. *The metrics do not behave as they would for general quality assessment*

A closer look at Tables VI-VIII reveals that the results of HD sequences are relatively better than those of CIF and Full-HD sequences. The reason is that the MVcopy and MVE methods conceal the errors considerably better than DMVE+BMA and AECOD for these HD sequences due to their 50 or 60 fps; therefore, the video clips are very different in quality which causes relatively more success for the metrics. However, we can see that the results of VSNR are disappointing for many HD and Full-HD sequences, while WMPDS acts more reliably for many sequences. The *MCDR*s of the CIF sequences in Table VI are very good for the *Bus* sequence (except for FLOSIM metric), but it is not the case for the *Silent* and *Foreman* sequences. VLAF performs much better than the others for *Tractor* and it is rather successful for other sequences, but it is not the case for *Foreman* and *Stefan*. Therefore, the concrete result is that, even though we usually see more correct judgments from WMPDS and VMAF, or the relative success of the metrics for *Bus* CIF, HD sequences, generally they are not trustable for comparing the error-concealed videos. The *MCDR* of 70% means the metric has a nearly 30% error rate in comparing the actual qualities.

Another important observation is the performance of PSNR. It is generally believed in the literature that PSNR is not as good as the other metrics for general video quality measurement. Some measures like SSIM, MOVIE, FLOSIM or others are considered better than PSNR for measurement of general quality of videos. But as we see in the above results, this is not the case for quality assessment of error-concealed videos, where PSNR's performance is better. There are many cases that PSNR provides higher *MCDR* than the newly developed metrics such as FAST method. The reason for the sometimes weaker performance of metrics such as SSIM in evaluating the quality of the error-concealed videos might be due to the different nature of error concealment distortion compared to the other types of distortions such as quantization, blurriness, blockiness and so on. The error concealment distortion is a kind of pixel displacement, and this distortion is localized in one region. Although it might be slightly displaced spatially through motion compensated coding, it is not spread all over the picture. This type of distortion might not be captured well by the single scale SSIM, since the structural similarity might not be much affected by the non-ideal error concealment. However, the pixel displacements can be better captured by the position based metrics, such as PSNR.

Even though VQA methods take into account the temporal distortion, it cannot be said that VQA methods perform better

Table IX The gain in *MCDR*s computed by equation (3) (Approach 2)

CIF sequences												
	Soccer		***Ice***		***Silent***		***Bus***		***Foreman***		***Stefan***	
SSIM	3.3		5.0		-5.1		5.1		-0.4		1.9	
PSNR	-3.1		5.0		3.5		2.3		1.1		3.8	
FSIM	-3.1		2.1		-4.8		2.3		-6.9		3.8	
FSIMc	-3.1		2.1		-4.8		2.3		-6.9		3.8	
GMSD	3.6		5.0		-0.9		-0.6		-1.3		4.8	
VSNR	-3.4		10.7		7.5		5.1		-1.1		6.2	
MS-SSIM	3.6		5.0		-3.1		-0.6		-5.3		4.8	
SEDLAI-V(I)	7.9		5.0		3.0		5.1		8.7		28.1	
STMAD(MAD)	16.1		2.1		7.5		5.1		-3.1		8.1	
HaarPSI	8.2		2.1		9.9		5.1		1.1		-6.2	
HD sequences												
	FourPeople		***KristenAnd Sara***		***Johnny***		***Vidyo1***		***Vidyo3***		***Mobcal***	
	LP1	LP2	LP1	LP2	LP1	LP2	LP1	LP2	LP1	LP2	LP1	LP2
SSIM	0.0	3.4	3.3	3.3	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
PSNR	0.0	0.0	0.0	3.3	0.0	0.0	0.0	0.0	0.0	3.3	3.7	0.0
FSIM	0.0	3.4	0.0	3.3	0.0	0.0	0.0	0.0	0.0	0.0	-3.7	-3.4
FSIMc	0.0	3.4	0.0	3.3	0.0	0.0	0.0	0.0	0.0	0.0	-3.7	-3.4
GMSD	0.0	0.0	0.0	3.3	0.0	0.0	0.0	0.0	0.0	3.3	-3.7	0.0
VSNR	-6.7	-31.0	-10.0	0.0	-6.9	0.0	-16.7	-16.7	-3.3	-3.3	-3.7	3.4
MS-SSIM	0.0	3.4	0.0	3.3	0.0	0.0	0.0	0.0	0.0	3.3	3.7	0.0
SEDLAI-V(I)	0.0	6.9	6.7	3.3	-3.4	-3.6	0.0	3.3	3.3	0.0	0.0	3.4
STMAD(MAD)	0.0	3.4	-3.3	3.3	0.0	7.1	4.2	0.0	0.0	6.7	-3.7	-3.4
HaarPSI	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	-3.7	-6.9
Full-HD sequences												
	Tractor		***Pedestrian _area***		***Crowd_run***		***Touchdown _pass***		***Rush_field_cuts***		***Speed_bag***	
	LP1	LP2	LP1	LP2	LP1	LP2	LP1	LP2	LP1	LP2	LP1	LP2
SSIM	2.7	0.0	4.9	0.0	3.7	-6.7	-2.1	2.0	-11.4	2.2	-2.0	-2.3
PSNR	2.7	-1.4	9.8	2.7	0.0	-10.0	6.4	0.0	2.3	21.7	0.0	18.6
FSIM	0.0	-1.4	-2.4	2.7	0.0	-10.0	2.1	-6.1	-4.5	32.6	-8.0	7.0
FSIMc	0.0	-1.4	-2.4	2.7	0.0	-10.0	2.1	-6.1	-4.5	32.6	-8.0	7.0
GMSD	1.4	-2.9	0.0	-2.7	0.0	-13.3	17.0	6.1	-9.1	19.6	0.0	11.6
VSNR	2.7	0.0	4.9	8.1	0.0	-3.3	10.6	0.0	2.3	10.9	-8.0	7.0
MS-SSIM	0.0	-1.4	12.2	5.4	0.0	-13.3	-8.5	10.2	2.3	28.3	-4.0	16.3
SEDLAI-V(I)	4.1	5.7	29.3	29.7	0.0	0.0	-	-	-	-	-	-
STMAD(MAD)	0.0	-2.9	12.2	8.1	0.0	-3.3	-8.5	-2.0	4.5	17.4	2.0	4.7
HaarPSI	2.7	0.0	0.0	-5.4	0.0	-6.7	12.8	-6.1	6.8	15.2	2.0	9.3

than IQA metrics, as tables VI-VIII show. The reason is that the temporal distortion generated by error propagation of error concealment is different from what is commonly seen in the datasets, so the features of the existing VQA methods are not sufficiently sensitive to this type of temporal distortion. In addition to the severity of the spatial distortion, how many frames are affected is also important. Another artifact generated by error propagation of error concealment is saturating the pixel values; i.e., it may become below zero or above 255. This is caused by the residual signals of the correctly received blocks which are added to the erroneous reference frame blocks; which in turn lead to out-of-range pixel values.

VII. Conclusion and Future Works

I/VQA metrics are mainly used to measure the video quality contaminated by environmental noise, compression distortion, transmission distortion, blurriness and so on. However, the application of these metrics to compare the quality of error concealment techniques had not been studied. The reason is that the distortions caused by error concealment techniques have not been reflected in the developed datasets. In this work, with the goal of evaluating the metrics for predicting the performance of error concealment techniques, an appropriate dataset was generated and is available in ECVD [50].

We showed that subjectively evaluating the error-concealed frame only is not necessarily representative of the quality of the video clips. Therefore, in our subjective tests, the error-concealed video clip was subjectively evaluated. For the objective tests, the error-concealed frame as well as the whole video clip were evaluated by various well-known I/VQA metrics. It was revealed that, compared to conventional video quality assessment, the metrics used in this study were not always as successful, and also their relative performance was different. For example, for conventional video quality assessment, PSNR gives a lower performance coefficient than the more recently developed metrics, but this was not the case in our study for assessing error-concealed videos, and PSNR outperformed SSIM and some other metrics for several sequences. The metric WMPDS which works based on the

pooling of the weighted PSNRs of the frames was more successful in our tests.

As future work, we can suggest two directions, as follows:

First, researchers should use an appropriate dataset; e.g. what we have provided in ECVD [50]. Other existing datasets do not represent the error concealment distortions in the video. The error concealment distortion is actually composed of three types of distortions concurrently, a) pixels displacement due to non-ideal MV recovery; b) non-uniform quantization distortion due to missing the residual signals; and c) error propagation due to erroneous reconstruction of the reference frame. The residual signals are added to the erroneous reference pixels, so we may have color saturated areas if the pixel values get out of range. These three types of distortions make evaluation of the error-concealed videos with the existing I/VQA metrics challenging, since the metrics are not designed and/or optimized for error concealment distortion.

Second, the metrics components should be adequately sensitive to the corrupted features in the error-concealed videos. For example, as it was noted for PSNR, the pixel location sensitive features can be quality representatives, but they are not sufficient. The other feature can be the number of corrupted frames due to error propagation. For example, we may have 4 highly corrupted frames or 8 frames but with moderate corruption, which one is preferred by the subjects and predicted by the metrics? As already mentioned, the temporal distortion here has particular aspects and critical importance. These considerations have not been accounted for in the development of the existing metrics.

REFERENCES

[1] Y. Wang and Q. F. Zhu, "Error control and concealment for video communication: A review," *Proc. IEEE,* vol. 86, no. 5, pp. 974–997, May 1998.

[2] M. Usman, X. He, M. Xu and K. Lam, "Survey of Error Concealment Techniques: Research Directions and Open Issues", *IEEE Picture Coding Symposium,* pp: 233-238, June 2015

[3] M. Fleury, S. Moiron, M. Ghanbari, "Innovations in video error resilience and concealment," *Recent Patents Signal Process*. vol. 1, no 2, pp. 1-11, 2011.

[4] D. M. Chandler, "Seven challenges in image quality assessment: Past, present, and future research," *ISRN Signal Process.*, vol. 2013,Nov. 2013

[5] A. C. Bovik, "Automatic prediction of perceptual image and video quality," *Proc. IEEE*, vol. 101, no. 9, pp. 2008–2024, Sep. 2013.

[6] S. Winkler and P. Mohandas, "The Evolution of Video Quality Measurement: From PSNR to Hybrid Metrics," *IEEE Transactions on Broadcasting*, vol. 54, no. 3, pp. 660-668, 2008

[7] M. Vranjes, S. Rimac-Drlje, K. Grgic, "Review of objective video quality metrics and performance comparison using different databases," *Signal Processing: Image Communication*, vol. 28, no. 1, pp. 1-19, 2013

[8] S. Chikkerur, V. Sundaram, M. Reisslein, and L. J. Karam, "Objective Video Quality Assessment Methods: A Classification, Review, and Performance Comparison," *IEEE Transactions on Broadcasting*, vol. 57, no. 2, pp. 165-182, 2011

[9] K. Seshadrinathan, R. Soundararajan, A. Bovik, and L. Cormack, "Study of subjective and objective quality assessment of video," *IEEE Trans. Image Process.*, vol. 19, no. 6, pp. 1427–1441, Jun. 2010.

[10] S. J. Kim, C. B. Chae. J. S. Lee, "Subjective and objective quality assessment of videos in error-prone network environments," *Springer Multimed Tools Appl*, vol. 25, no. 12, pp. 6849–6870, 2016.

[11] A. K. Moorthy, K. Seshadrinathan, R. Soundararajan, and A. C. Bovik, "Wireless video quality assessment: A study of subjective scores and objective algorithms," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 20, no. 4, pp. 587–598, Apr. 2010

[12] S. Wan, F. Yan, Z. Xie, "Evaluation of video quality degradation due to packet loss," *Intelligent Signal Processing and Communication Systems (ISPACS)*, 2010.

[13] J. You, J. Korhonen, A. Perkis,"Spatial and temporal pooling of image quality metrics for perceptual video quality assessment on packet loss streams,"*In: Proc. IEEE int. conf. acoust., speech and sig. proc.*, vol 7491, pp 1002–1005.

[14] A. R. Reibman, V. A. Vaishampayan, and Y.Sermadevi, "Quality Monitoring of Video Over a Packet Network," *IEEE Transactions on Multimedia*, vol. 6, no. 2, pp. 327-334, April 2004

[15] A. K. Moorthy, L. K. Choi, A. C. Bovik, and G. deVeciana, "Video quality assessment on mobile devices: Subjective, behavioral and objective studies," *IEEE J. Sel. Topics Signal Process.*, vol. 6, no. 6, pp. 652–671, Oct. 2012

[16] D. Ghadiyaram, A. C. Bovik, H. Yeganeh, R. Kordasiewicz, and M. Gallant, "Study of the effects of stalling events on the quality of experience of mobile streaming videos," *in Proc. IEEE Global Conf. Signal Inf. Process*. 2014.

[17] Z. Duanmu, K. Zeng, K. Ma, A. Rehman, and Z. Wang, "A Quality-of-Experience Index for Streaming Video," *IEEE Journal of Selected Topics In Signal Processing*, vol. 11, no. 1, pp. 154-166, February 2017

[18] C. G. Bampis, Z. Li, A. K. Moorthy, I. Katsavounidis, A. Aaron and A. C. Bovik, "Study of Temporal Effects on Subjective Video Quality of Experience," *IEEE Transactions on Image Processing*, vol. 26, no. 11, pp. 5217-5231, 2017

[19] D. Ghadiyaram, J. Pan, and A. C. Bovik,"A Subjective and Objective Study of Stalling Events in Mobile Streaming Videos," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 29, no. 1, pp. 183–197, 2019

[20] Z. Wang and A. C. Bovik, "Mean squared error: Love it or leave it? A new look at signal fidelity measures," *IEEE Signal Process. Mag.*, vol. 26, no. 1, pp. 98–117, Jan. 2009

[21] B. Girod, "What's wrong with mean-squared error?" in *Digital Imagesand Human Vision*, A. B. Watson, Ed. Cambridge, MA: MIT Press,1993, pp. 207–220

[22] Q. Huynh-Thu and M. Ghanbari, "Scope of validity of PSNR in image/video quality assessment," *Electronics Letters*, vol. 44, pp. 800–801, 2008

[23] J. Korhonen, J. You, "Peak signal-to-noise ratio revisited: is simple beautiful?," *Fourth International Workshop on Quality of Multimedia Experience*, 2012.

[24] B.P. Bondzulic, B.Z. Pavlovic, V.S. Petrovic and M.S. Andric, "Performance of peak signal-to-noise ratio quality assessment in video streaming with packet losses," *Electronics Letters*, vol. 52, no. 6, pp. 454-456, 2016.

[25] Z. Wang, A. C. Bovik, H. R. Sheikh, E. P. Simoncelli, "Image Quality Assessment: From Error Visibility to Structural Similarity," *IEEE Transactions On Image Processing*, vol. 13, no. 4, pp. 600-612, April 2004.

[26] Z. Wang, E. P. Simoncelli and A. C. Bovik, "Multi-scale structural similarity for image quality assessment," *IEEE Asilomar Conference Signals, Systems and Computers*, November 2003.

[27] L. Zhang, L. Zhang, X. Mou, and D. Zhang, "FSIM: A Feature Similarity Index for Image Quality Assessment," *IEEE Transactions on Image Processing*, vol. 20, no. 8, pp. 2378–2386, 2011

[28] W. Xue, L. Zhang, X. Mou, and A. C. Bovik, "Gradient Magnitude Similarity Deviation: A Highly Efficient Perceptual Image Quality Index," *IEEE Transactions on Image Processing*, vol. 23, no. 2, pp. 684-695, 2014.

[29] D. M. Chandler and S. S. Hemami, "VSNR: A wavelet-based visual signal-to-noise ratio for natural images," *IEEE Trans. Image Process.*, vol. 16, no. 9, pp. 2284–2298, 2007

[30] Songnan Li, Fan Zhang, Lin Ma, King Ngi Ngan, "Image quality assessment by separately evaluating detail losses and additive impairments", *IEEE Transactions on Multimedia*, vol. 13, no. 5, pp. 935-949, Oct. 2011

[31] E. C. Larson and D. M. Chandler, "Most apparent distortion: Full-reference image quality assessment and the role of strategy," *Journal of Electronic Imaging*, vol. 19, no. 1, Article ID 011006, 2010

[32] M. H. Pinson, S. Wolf, "A New Standardized Method for Objectively Measuring Video Quality," *IEEE Transactions on Broadcasting*, vol. 50, no. 3, pp. 312-322, Sept. 2004

[33] K. Seshadrinathan, A. C. Bovik, "Motion tuned spatio-temporal quality assessment of natural videos," *IEEE Trans. Image Process.*, vol 19. no. 2, pp. 335-350, Feb. 2010

[34] K. Manasa and S. S. Channappayya, "An optical flow-based full reference video quality assessment algorithm," *IEEE Transactions on Image Processing*, vol. 25, no. 6, pp. 2480–2492, 2016

[35] S. Li, L. Ma, and K. N. Ngan, "Video quality assessment by decoupling additive impairments and detail losses," in *Proc. 3rd Int. Workshop Qual. Multimedia Experience*, 2011, pp. 90–95

[36] S. Li, L. Ma, and K. N. Ngan,"Full-Reference Video Quality Assessment by Decoupling Detail Losses and Additive Impairments," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 22, no. 7, pp. 1100-1112, 2012

[37] P. Vu, C. Vu, and D. Chandler, "A spatiotemporal most apparent distortion model for video quality assessment," in *Proceedings of the 18th IEEE International Conference on Image Processing (ICIP '11)*, pp. 2505–2508, September 2011

[38] T. Liu, Y. Wang, J. M. Boyce, H. Yang, and Z. Wu, "A Novel Video Quality Metric for Low Bit-Rate Video Considering Both Coding and Packet-Loss Artifacts," *IEEE Journal of Selected Topics In Signal Processing*, vol. 3, no. 2, pp. 280-293, April 2009

[39] V. Seferidis, M. Ghanbari, D. E. Pearson, "Forgiveness effect in the subjective assessment of packet video," *Electronics Letters*, 28:21, pp. 2013-2014, October 1992

[40] S. Bosse, D. Maniry, K. R. Müller, T. Wiegand, W. Samek, "Deep Neural Networks for No-Reference and Full-Reference Image Quality Assessment," *IEEE Transactions on Image Processing*, vol. 27, no. 1, pp. 206-219, January 2018

[41] ITU-T Recommendation, P.910, "Subjective video quality assessment methods for multimedia applications," *International Telecommunication Union*, Tech. Rep., 2008

[42] ITU-T Recommendation, P.913, "Methods for the subjective assessment of video quality, audio quality and audiovisual quality of Internet video and distribution quality television in any environment International Telecommunication Union, Tech. Rep., 2016.

[43] ITU-T J.149, "Method for specifying accuracy and cross-calibration of Video Quality Metrics (VQM)," *International Telecommunication Union*, Mar. 2004

[44] P. Hanhart, L. Krasula, P. L. Callet, and T. Ebrahimi, "How to Benchmark Objective Quality Metrics from Paired Comparison Data?," *International Conference on Quality of Multimedia Experience (QoMEX)*, 2016.

[45] F. De Simone *et al.*, "EPFL-PoliMI video quality assessment database," 2009 [Online]. Available: http://vqa.como.polimi.it/

[46] F. Boulos, W. Chen, B. Parrein, and P. Le Callet, "IRCCyN IVC SD RoI database," 2009 [Online]. Available: http://www.irccyn.ec-nantes.fr/spip.php?article551

[47] F. Zhang, S. Li, L. Ma, Y. C. Wong, and K. N. Ngan, "IVP subjective quality video database," 2011 [Online]. Available: http://ivp.ee.cuhk.edu.hk/research/database/subjective

[48] K. Seshadrinathan, R. Soundararajan, A. C. Bovik, and L. K. Cormack, "LIVE video quality database," 2010 [Online]. Available: http://live.ece.utexas.edu/research/quality/live_video.html

[49] Y. Wang *et al.*, "Poly@NYU video quality databases," 2008 [Online]. Available: http://vision.poly.edu/index.html/index.php?n=HomePage.QualityAssessmentDatabase

[50] M. Kazemi, M. Ghanbari, S. Shirmohammadi, "Error-Concealed Video Dataset (ECVD)", IEEE Dataport, 2020. [Online]. Available: http://dx.doi.org/10.21227/31wz-b576

[51] Q. Peng, T. Yang, and C. Zhu, "Block-based temporal error concealment for video packet using motion vector extrapolation," *in IEEE Int. Conf. Commun., Circuits Syst. West Sino Expo.*, Jul. 2002, vol. 1, pp. 10–14.

[52] M. C. Hwang, J. H. Kim, D. T. Duong, and S. J. Ko, "Hybrid temporal error concealment methods for block-based compressed video transmission," *IEEE Trans. Broadcast.*, vol. 54, no. 2, pp. 198–207, Jun. 2008

[53] X. Qian, G. Liu, and H. Wang, "Recovering connected error region based on adaptive error concealment order determination," *IEEE Trans. Multimedia*, vol. 11, no. 4, pp. 683–695, Jun. 2009.

[54] K. Seshadrinathan and A. C. Bovik, "Temporal hysteresis model of time varying subjective video quality," *in Proc. IEEE Int. Conf. Acoust., Speech Signal Process*., May 2011, pp. 1153–1156.

[55] J. Park, K. Seshadrinathan, S. Lee, and A. C. Bovik, "Video quality pooling adaptive to perceptual distortion severity," *IEEE Trans. Image Process.*, vol. 22, no. 2, pp. 610–620, Feb. 2013.

[56] S. Katsigiannis, J. Scovell, N. Ramzan, L. Janowski, P. Corriveau, M. A. Saad, G. V. Wallendael, "Interpreting MOS scores, when can users see a difference? Understanding user experience differences for photo quality," *Quality and User Experience*, vol. 3, no. 6, 2018.

[57] R. Reisenhofer, S. Bosse, G. Kutyniok, T. Wiegand, "A Haar wavelet-based perceptual similarity index for image quality assessment," *Signal Processing: image communication*, vol. 61, pp. 33-43, 2018.

[58] M. A. Papadopoulos, A. V. Katsenou, D. Agrafiotis, D. R. Bull, "A multi-metric approach for block-level video quality assessment," *Signal Processing: image communication*, vol. 78, pp. 152-158, 2019.

[59] J. Kim, H. Zeng, D. Ghadiyaram, S. Lee, L. Zhang, and A. C. Bovik, "Deep convolutional neural models for picture-quality prediction: Challenges and solutions to data-driven image quality assessment," *IEEE Signal Process. Mag.*, vol. 34, no. 6, pp. 130–141, Nov. 2017.

[60] T. R. Goodall and A. C. Bovik, "Detecting and Mapping Video Impairments," *IEEE Transactions on Image Processing*, vol. 28, no. 6, pp. 2680 - 2691 ,2019.

[61] L. K. Choi, A. C. Bovik, "Video quality assessment accounting for temporal visual masking of local flicker," *Signal Processing: image communication*, vol. 67, pp. 182-198, 2018.

[62] J. Wu, Y. Liu, W. Dong, G. Shi, and W. Lin, "Quality Assessment for Video with Degradation Along Salient Trajectories," *IEEE Transactions on Multimedia*, vol. 21, no. 11, 2738 – 2749, 2019.

[63] Zhi Li, Anne Aaron et al, "Toward A Practical Perceptual Video Quality Metric," *Netflix TechBlog*, June, 2016

[64] X. H. Van, and B. Jeon, "Joint Layer Prediction for Improving SHVC Compression Performance and Error Concealment," *IEEE Transactions on Broadcasting*, vol. 65, no. 3, pp. 504-520, 2019.

[65] P. C. Huang, J. R. Lin, G. L. Li, K. H. Tai, and M. J. Chen, "Improved Depth-Assisted Error Concealment Algorithm for 3D Video Transmission," *IEEE Transactions on Multimedia*, vol. 19, no. 11, pp. 2625 – 2632, 2017

[66] M. Usman, X. He, K. M. Lam, M. Xu, S. M. M. Bokhari, J. Chen, and M. A. Jan, "Error Concealment for Cloud–based and Scalable Video Coding of HD Videos," *IEEE Transactions on Cloud Computing*, vol. 7, no. 4, pp. 975 – 987, 2019

[67] B. Chung, and C. Yim, "Bi-sequential Video Error Concealment Method Using Adaptive Homography-based Registration," *IEEE Transactions on Circuits and Systems for Video Technology (Early Access)*, doi: 10.1109/TCSVT.2019.2909564

[68] J. F. M. Carreira, P. A. Assunção, S. de Faria, E. Ekmekcioglu, A. Kondoz, "Error Concealment-Aware Encoding for Robust Video Transmission," *IEEE Transactions on Broadcasting*, vol. 65, no. 2, pp. 282-293, 2018.

[69] R Aldridge, J Davidoff, M Ghanbari, D Hands, D Pearson, "Recency effect in the subjective assessment of digitally-coded television pictures," *Fifth International Conference on Image Processing and its Applications, 4-6 July, Edinburgh, UK*, 1995.

Mohammad Kazemi received the B.Sc. degree from the Isfahan University of Technology, Isfahan, Iran, in 2003, and the M.Sc. and PhD degrees from the Sharif University of Technology (SUT), Tehran, Iran, both in electrical engineering, in 2005 and 2012, respectively. He currently works with Electrical Engineering Department, University of Isfahan, Iran, as an Assistant Professor. His current research interests include the areas of image/video processing, coding, and video transmission over lossy channels. He is also interested in digital systems design, and applications of Artificial Intelligence for image/video content analysis.

Mohammad Ghanbari (IEEE M'78–SM'97–F'01–LF'14) is a Professor at the School of Electrical and Computer Engineering, University of Tehran, as well as an Emeritus Professor at the School of Computer Science and Electronic Engineering, University of Essex, U.K. He has authored or co-authored eight books and has registered for 13 international patents and authored over 750 technical papers on various aspects of video networking, many of which have had fundamental influences in this field. These include video/image compression, layered/scalable video coding, video over networks, video transcoding, motion estimation, and video quality metrics. He is internationally best known for pioneering work on layered video coding, for which he received the IEEE Fellowship in 2001 and was promoted as an IEEE Life Fellow in 2014. His book Video Coding: An Introduction to Standard Codecs (IET Press, 1999) received the Rayleigh Prize as the best book of the year 2000 by IET. He has also received several prizes, such as Reeves Prize for best paper award in 1995 and 14th Khwarizmi international award for work on video networking in 2001.

Shervin Shirmohammadi (M '04, SM '04, F '17) received his Ph.D. in Electrical Engineering from the University of Ottawa, Canada, where he is currently a Professor with the School of Electrical Engineering and Computer Science. He is Director of the Distributed and Collaborative Virtual Environment Research Laboratory, doing research in Applied AI for multimedia systems and networks, specifically video systems, gaming systems, and multimedia-assisted healthcare systems. The results of his research, funded by more than $14 million from public and private sectors, have led to over 350 publications, 3 Best Paper awards, over 70 researchers trained at the postdoctoral, PhD, and Master's levels, over 20 patents and technology transfers to the private sector, and a number of awards. He is the Editor-in-Chief of IEEE Transactions on Instrumentation and Measurement, and an Associate Editor of ACM Transactions on Multimedia Computing Communications and Applications, having been numerously recognized as the Associate Editor of the Year by both of these and other journals.

Dr. Shirmohammadi is an IEEE Fellow *for contributions to multimedia systems and network measurements*, winner of the 2019 George S. Glinski Award for Excellence in Research, a Lifetime Senior Member of the ACM, a University of Ottawa Gold Medalist, and a licensed Professional Engineer in Ontario