Study of the Subjective Visibility of Packet Loss Artifacts in Decoded Video Sequences

Korhonen, Jari

Published in:
IEEE Transactions on Broadcasting

Link to article, DOI:
10.1109/TBC.2018.2832465

Publication date:
2018

Document Version
Peer reviewed version

Citation (APA):
Korhonen, J. (2018). Study of the Subjective Visibility of Packet Loss Artifacts in Decoded Video Sequences. IEEE Transactions on Broadcasting. DOI: 10.1109/TBC.2018.2832465

Study of the Subjective Visibility of Packet Loss Artifacts in Decoded Video Sequences

Jari Korhonen, *Member, IEEE*

***Abstract*—Packet loss is a significant cause of visual impairments in video broadcasting over packet-switched networks. There are several subjective and objective video quality assessment methods focused on the overall perception of video quality. However, less attention has been paid on the visibility of packet loss artifacts appearing in spatially and temporally limited regions of a video sequence. In this paper, we present the results of a subjective study, using a methodology where a video sequence is displayed on a touchscreen and the users tap it in the positions where they observe artifacts. We also analyze the objective features derived from those artifacts, and propose different models for combining those features into an objective metric for assessing the noticeability of the artifacts. The practical results show that the proposed metric predicts visibility of packet loss impairments with a reasonable accuracy. The proposed method can be applied for developing packetization and error recovery schemes to minimize the subjectively experienced distortion in error-prone networked video systems.**

***Index Terms*—Digital television broadcasting, H.264, IPTV, multimedia broadcasting, QoE, video processing.**

I. Introduction

THE PERCEIVED quality of video sequences transmitted over communications networks is often impaired by several different types of artifacts, most notably compression noise and channel artifacts caused by transmission errors, such as packet losses. In many applications, such as optimization of coding and transmission parameters for adaptive streaming, it would be desirable to estimate the subjective quality in real time without human involvement, using a dedicated algorithm, usually referred to as an objective quality metric. To develop and validate objective metrics, subjective quality ratings are required as ground truth. In order to obtain this ground truth, subjective quality assessment studies need to be organized, involving human test subjects to assess the perceived video quality.

Several different methodologies for subjective quality assessment have been proposed in the prior art. Most common methodologies use some kind of rating scale, either Absolute Category Rating (ACR), such as the five point scale ranging from "bad" to "excellent", or relative impairment scale where distortions are assessed in comparison to a reference sequence, using a scale ranging for instance from "impairments are imperceptible" to "impairments are very annoying" [1], [2]. The obtained ratings are then converted to a numerical scale, and then the Mean Opinion Score (MOS) can be computed, representing the subjective quality of the video. Alternative methodologies comprise, e.g., pairwise comparisons [3], [4] and rank ordering [5], [6].

Subjective quality assessment based on ratings, pairwise comparisons or rank ordering are useful for assessing the overall quality of short video sequences. Unfortunately, they cannot be used for assessing the relative annoyance of time-varying individual artifacts, such as quality fluctuations caused by rate adaptation or glitches caused by transmission errors, such as packet losses in packet-switched networks. To capture the dynamics of temporally varying video quality, continuous quality assessment techniques have been proposed. Those methodologies use, e.g., a continuously adjusted slider or knob to dynamically indicate quality changes [7], [8], or a button that is pressed when quality drops to unacceptable level, or a glitch is observed [9]–[11].

Unfortunately, those continuous quality assessment methods are still restricted to the temporal dimension, and the spatial location of the artifacts is not determined by the test subjects. As High Definition (HD) resolution is becoming a commonplace, it is not rare that two or more separate packet loss artifacts overlap in the temporal dimension within the same sequence. To evaluate the relative visibility of artifacts occupying different locations in both spatial and temporal dimensions, we have proposed a novel methodology, where test subjects indicate the location of observed artifacts by tapping a touchscreen in the respective position [12]. The percentage of users who have observed the artifact can then be used as a measure for the subjective visibility of that artifact.

In this paper, we present an in-depth analysis of the relationship between the subjective visibility and the objective characteristics of individual packet loss impairments, based on the subjective data obtained in our earlier study [12]. We also propose a novel Full-Reference (FR) objective model for estimating the noticeability of individual packet loss artifacts, based on the features derived from their objective characteristics. The insights from this study can be applied for estimating user satisfaction for video streaming and broadcasting services under different channel error conditions, as well as development and optimization of coding and transmission

Manuscript received November 28, 2017; revised April 10, 2018; accepted April 18, 2018. This work was supported by the National Science Foundation of China under Grant 61772348.

The author was with the Department of Photonics Engineering, Technical University of Denmark, 2800 Kongens Lyngby, Denmark. He is now with Shenzhen University, College of Computer Science and Software Engineering, Shenzhen 518060, China (e-mail: jari.t.korhonen@ieee.org).

Color versions of one or more of the figures in this paper are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/TBC.2018.2832465

schemes for networked video systems. The proposed FR packet loss visibility model can also be used for benchmarking the No-Reference (NR) methods for assessing packet loss visibility.

The rest of the paper is organized as follows. In Section II, we discuss the background and the related work, including typical characteristics of packet loss impairments, as well as the relevant subjective and objective quality assessment methods concerning packet loss artifacts. In Section III, we summarize the subjective experiment for collecting the data used in this paper [12], propose a novel method for assessing macroblock level error visibility, explain a revised method to determine error clusters, and then show how the error clusters are assigned to taps on the touchscreen. In Section IV, we propose different objective features characterizing packet loss artifacts and analyze their correlation with subjective visibility of packet loss artifacts. We also propose a novel model combining different features to predict the subjective visibility more precisely. Finally, conclusions are given in Section V.

II. Background and Related Work

Packet losses occur in communication systems where the underlying transmission protocols do not guarantee reliable delivery of transport packets. For the networks based on Internet Protocol (IP), this is the case if User Datagram Protocol (UDP) is used as a transport protocol. Even though video streaming based on the reliable Transmission Control Protocol (TCP) is becoming more common, there are still applications where video communications deploying UDP is a reasonable option. For example, in IP-based cable TV networks, IP multicast is typically deployed, which prevents the use of TCP. In addition, video transmission over TCP is prone to frame freezes, and especially for interactive applications, packet loss artifacts may be preferred to frame freezes.

A. Characteristics of Packet Loss Artifacts

Typical visual characteristics of packet loss impairments are fundamentally different from source distortion [13], [14]. The severity of channel distortion depends on many different factors: packet loss rate and pattern plays a major role, but the video content and the coding parameters also have a high impact on the actual distortion as experienced by the end user [13]. By using efficient error concealment, the visual quality distortion can be significantly alleviated. In the best case, the remaining distortion may be nearly invisible to a human observer.

Traditionally, error concealment techniques are classified as spatial methods, where the lost areas are reconstructed by interpolating from the spatially surrounding pixels, and temporal methods, where lost areas are copied from temporally adjacent frames. Several sophisticated methods for both spatial and temporal error concealment have been proposed in [14]–[17]. Unfortunately, the most efficient methods are relatively complex in terms of computation and implementation [17]. This is why practical video decoders usually rely on simpler methods, such as bilinear interpolation for spatial error concealment and motion copy for temporal error concealment [18]. These methods are able to produce satisfactory results for typical video contents with a relatively small packet loss burden.

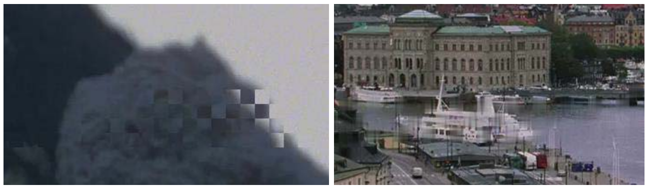

a) Temporal concealment b) Spatial concealment

Fig. 1. Examples of temporal and spatial error concealment with FMO. Temporal concealment (a) appears as misplaced blocks; spatial concealment (b) appears as blurry blocks, due to interpolation.

The performance of error concealment can be further boosted by employing codec specific error resilience features, such as Flexible Macroblock Ordering (FMO) in H.264/AVC [19]. With FMO, it is possible to interleave consequent macroblocks in different transmission units so that in case of an individual packet loss, there are adjacent macroblocks available next to each lost macroblock, facilitating interpolation-based concealment of lost macroblocks [19]. Practical experiments have shown that FMO can improve video quality in the presence of packet losses, at the cost of a slightly decreased compression efficiency [19], [20].

Two main types of packet loss artifacts can be identified, related to different error concealment techniques [13], [14]. Spatial interpolation is typically used when there is no reference frame available (e.g., the first frame in the group of pictures or the frame immediately after a scene change), or when the spatial activity level in the affected region is low (e.g., blue sky or some other uniform surface). Spatial concealment methods are based on interpolation from the pixels in the correctly received regions, often resulting in blur effect. Another class of packet loss artifacts is attributed to temporal error concealment methods, aiming to replace the lost blocks by copying the best matching block from the previous frame. Temporal error concealment usually works well on static background, but often causes severe distortion in regions with intensive motion. In this case, the missing blocks tend to be misplaced in respect with the neighboring blocks. An example of artifacts resulting from both temporal and spatial error concealment is shown in Fig. 1.

Even when the location of the lost macroblocks is available to the decoder, it is difficult to know to which extent other macroblocks are affected by the losses. This is because practical encoded video sequences typically use temporal prediction, leading to error propagation between macroblocks. Division of macroblocks into smaller prediction blocks can make the prediction structure rather complex and can cause even more apparent temporal artifacts.

An example of error propagation related to temporal error concealment is shown in Fig. 2. The impact of error propagation disappears when a new Group of Pictures (GOP) starts with a self-contained I-frame. In video sequences with intensive motion, the impact of error propagation may also vanish

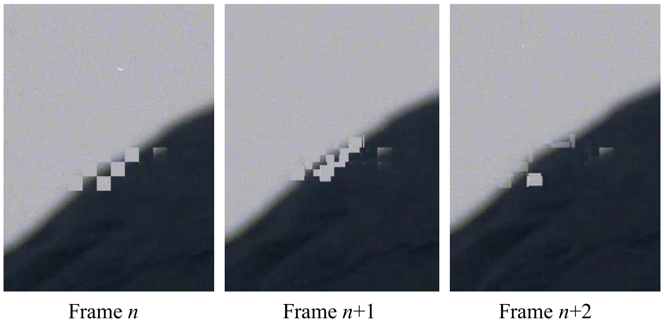


Fig. 2. Error propagation illustrated. Due to temporal prediction, checkerboard shaped artifact in Frame *n* is scattered into an irregularly shaped artifact in Frames $n + 1$ and $n + 2$.

gradually, as the damaged pixels are eventually fully replaced by the decoded prediction residuals. It is very difficult, if not impossible, to estimate the visual impact of error propagation analytically. In the related work, some models have been proposed for quality degradation that is related to error propagation [21]–[23], but those models can predict distortion with a statistically acceptable accuracy only on the sequence level, not on the frame level.

B. Subjective Assessment of Packet Loss Artifacts

Several subjective assessment studies were conducted to evaluate packet loss artifacts [24]–[27]. There are also some publicly available databases with video sequences distorted with simulated packet losses. Probably the best known database focusing exclusively on packet loss artifacts is the EPFL-PoliMi database, that contains videos in CIF (352×288 pixels) and 4CIF (704×576 pixels) resolutions, with packet loss rates ranging from 0.1% to 10% [26]. Another well-known database, including both coding and packet loss artifacts, is the LIVE database [28]. Public subjective video quality databases are usually annotated with the MOS values for each test sequence, possibly along with the standard deviation or even the scores given by individual test subjects. It is also possible to assess Just Noticeable Distortion (JND) points [29] instead of MOS; however, we are not aware of any JND-based databases focusing on packet loss artifacts.

Unfortunately, MOS is an estimate of the overall quality, and from plain MOS it is not possible to extract information about the individual artifacts and factors contributing to the overall score. To obtain more detailed information on the temporal dynamics related to packet loss artifacts and temporal quality variation in general, several continuous quality assessment methods have been proposed. For instance, the Single Stimulus Continuous Quality Evaluation (SSCQE) is a standardized method in which video quality is assessed by moving a slider simultaneously as the perceived quality changes [7]. The SSCQE method has been successfully employed for assessing video sequences with packet loss impairments [24], [25]. Borowiak *et al.* have proposed a continuous method where decreased audiovisual quality can be compensated by turning a knob [8]. However, to the best of our knowledge, this methodology has not been used to evaluate packet loss impairments.

Reibman and Poole [10] proposed a method where a space bar is pressed when a packet loss artifact is observed [30], [31]. Each packet loss is then characterized by a subjective visibility index, computed as the percentage of users who have observed the loss. In a similar fashion, Argyropoulos *et al.* [11] conducted an experiment using a button to indicate observed packet losses; in addition, the test sequences were also rated using an 11-point ACR scale. Jumisko-Pyykkö *et al.* [9] have also used a button to indicate unacceptable distortion caused by packet losses on audiovisual content.

The main weakness of the abovementioned continuous quality assessment methods is that they only give information about the presence of distortions in the temporal dimension. However, as high resolution consumer video streams are becoming more common, each encoded video frame needs to be fragmented in a large number of transport units. This is why a common scenario in high resolution video streaming may involve several temporally overlapping packet loss impairments occupying different spatial locations. On a large display, attention can be drawn to the major artifact, and simultaneous minor impairments are possibly not noticed. To obtain information about the spatial location of the observed impairments, we have proposed a methodology where the video is played on a touchscreen, and the user taps it in the position where a packet loss artifact is observed [12]. We will discuss the experiences gained from a subjective study employing the methodology in more details in Section III.

C. Objective Assessment of Packet Loss Artifacts

Typically, the design goal for video quality metrics is to provide good performance across different types of distortions, compression noise and transmission errors in particular. All the well-known objective Full-Reference (FR) video quality metrics, such as Peak Signal-to-Noise Ratio (PSNR), Structural Similarity Index Metric (SSIM) [32], Video Quality Metric (VQM) [33] and MOVIE [34] have been extensively tested with video sequences including also packet loss artifacts [30], [35]–[38]. There is evidence that many objective FR metrics are not capable to predict the perceptual impact of compression and channel artifacts equally well [35]–[37]; however, the best metrics usually achieve acceptable results in scenarios where different artifacts are present [35], [38].

The objective metrics discussed above are FR metrics, which means that they use the decoded video signal with impairments as input, along with the non-impaired reference video signal. However, in practical use cases of video streaming, packet losses are experienced by end users, who do not have access to the reference video. Therefore, No-Reference (NR) metrics are important for real-time quality monitoring in streaming applications. Since NR metrics use only the impaired video as input, they need to evaluate distortions indirectly by detecting features that are usually related to compression or channel noise, such as blockiness and blurriness [39].

It is challenging to distinguish features representing impairments from features that are part of the source content, and this is why NR metrics in general are not capable of predicting subjective video quality as accurately as state-of-the-art FR metrics. It has been shown that different distortion specific metrics (e.g., blur, blockiness and jerkiness) do not work well alone for NR assessment of global video quality, but they can be combined into a more accurate generic metric by machine learning [40]–[42]. Unfortunately, video quality metrics based on machine learning tend to be prone to irreproducibility, due to overfitting and the fact that the commonly used machine learning algorithms do not include a mechanism to ensure that different features are combined in a consistent manner [42]. To make sure that a learning-based quality metric considers packet loss artifacts properly, the metric has to be trained using video sequences with similar artifacts.

From the perspective of our study, the related work about NR packet loss artifact detection is especially interesting. Montard and Bretillon compute standard deviations of pixel intensities for each macroblock to detect suspicious macroblocks. As secondary criteria, they assess the similarity of the suspected macroblock and its neighbors to decide if the macroblock is considered lost [43]. Rui *et al.* [44] use edge detection along the macroblock borders to detect blocky packet loss artifacts. Ong *et al.* [45] also measure blockiness to detect packet loss artifacts; in addition, they measure inter-frame similarity to detect frame freezes and losses. In a similar fashion, Teslic *et al.* also measure gradients to detect edges at macroblock borders, and use this information to detect suspected packet loss artifacts [46].

Shabtay *et al.* [47] split the error detection task in two parts: detection of temporal error concealment, and detection of spatial error concealment. To detect temporal error concealment, Mean Absolute Difference (MAD) between each macroblock and the respective macroblock in the previous frame is computed. The authors assume that if MAD is small, the macroblock may have been directly copied from the previous frame. If an edge is detected at the upper or lower macroblock border, the suspicious macroblock is considered temporally concealed. For spatial error concealment, the authors assume that the lost macroblocks are interpolated using the edge pixels above and below the lost macroblock. This kind of error concealment makes vertically aligned structures smoother, which can be detected by comparing the gradients measured in vertical and horizontal directions.

Valenzise *et al.* also consider both temporal and spatial error concealment in their method [13]. In contrast to [47], motion compensation is also used to detect temporal error concealment. MAD between the concerned macroblock and the motion compensated reference macroblock, together with the variance of the local motion vectors computed from the decoded video, are used as discriminative features to indicate temporal error concealment. For detecting spatial error concealment, the authors assume that the used interpolation kernel is known, and therefore each macroblock can be compared directly against its spatially concealed version. The authors also propose a method for estimating the most likely positions of corrupted macroblocks, after first estimating the probabilities for individual macroblocks to be corrupted.

Detecting packet loss artifacts directly from decoded video is a challenging problem, since different forms of error concealment lead to artifacts with very different statistical characteristics [48]. Even though several methods have been proposed in the prior art, they tend to rely on some simplifying assumptions about the used concealment method. The best methods can detect an appearing packet loss artifact reasonably accurately. However, the shape and extent of the artifact is often evolving along time, due to temporal prediction and error propagation. As the duration of a packet loss artifact has a great impact on its visibility, more research is needed on the temporal dynamics of packet loss impairments.

III. Assessing Packet Loss Impairments

To assess the visibility of individual error clusters, we conducted a subjective study, where the test video is displayed on a touchscreen and the task for the test subjects is to tap the screen where they notice an appearance of a packet loss artifact. The subjective experiment and the preliminary results were first published in [12]. In this section, we summarize the study, propose a novel method for assessing the visibility of impairments at the macroblock level, and finally present revised algorithm for determining the error clusters and assigning each tap to an error cluster.

A. Subjective Study

For the subjective study, we used source content obtained from Consumer Digital Video Library (http://www.cdvl.org) [49]. The source video is about six minutes long, containing different scenes with diverse spatial and temporal characterstics. The original resolution and format was Full HD (1920×1080 pixels), 50 frames per second, in YUV4:4:4 coding. To facilitate processing, we converted the original video to YUV4:2:0 format, 25 frames per second, and encoded it with H.264/AVC reference codec (version 12.4) [50], using GOP length of 25 frames and FMO enabled. The audio track was not used. To guarantee stable source distortion level, Quantization Parameter (QP) was fixed to 24. To allow high granularity of packet loss impairments, each frame was divided in 42 slices, i.e., 200 macroblocks per Network Adaptation Layer Unit (NALU), one NALU per packet. To facilitate error concealment, FMO was enabled with the chessboard pattern.

For packet loss simulation, we used a simple MATLAB script to drop packets randomly. In order to produce meaningful output for the subjective experiment, we used different packet loss rates for different sections of the video: since packet loss impairments are typically more visible in regions with high spatial activity level, we used lower packet loss rate in those regions to avoid an overwhelming amount of distortions. The average packet loss rate was approximately 1.5%. It should be noted that our intention was not to simulate a realistic network scenario, where packet losses tend to appear in a bursty and sporadic manner. We assume that by producing more frequently distributed impairments, we can reduce the length of the experiment and obtain more informative results by keeping the test subjects more focused and motivated.

For error concealment, we used the standard techniques implemented in H.264 reference codec [50], as explained in [18]. For I-frames, spatial error concealment based on weighted interpolation is used. For P- and B-frames, temporal error concealment is used. The temporal error concealment scheme attempts to predict the motion vectors from the correctly received macroblocks. If the average motion is below a predefined threshold, missing macroblocks are just copied from the same position in the previous frame [18]. Even though the method is rather simple, it usually gives satisfactory results, and most of the real-life video broadcasting systems still rely on similar error concealment techniques.

The subjective test method is straightforward: the test subjects were instructed to view the video displayed on a touchscreen and to tap the screen when they observe a packet loss artifact, in the position where the artifact appeared. For playing the video and recording the taps, we developed a test program for this specific purpose, using C++ and Qt Creator platform. As hardware, we used Dell's panel PC (Windows 7) with 21.5 inch touchscreen. Before each session, a brief introduction was given, including examples of packet loss artifacts. Twenty test subjects participated in the study: 7 females, 13 males, from 20 to 33 years old. According to the feedback from the test subjects, the task was considered challenging but interesting.

B. Macroblock Level Error Visibility

In [12], we used the Mean Squared Error (MSE) between the impaired and non-impaired macroblocks (luma component only) as a macroblock level indicator of error visibility. Unfortunately, MSE does not predict the visual distortion very accurately in all situations. For example, if a macroblock containing a detailed texture is displaced by one pixel (a common artifact caused by temporal error concealment), MSE is typically very large. However, in this case, a human observer can hardly see any visual impairment at all. On the other hand, if a macroblock is located on a smooth uniform surface, even a small change in the tone can lead to a noticeable impairment, if the neighboring macroblock remains the same.

To overcome the limitations of MSE, we have proposed a more appropriate method for estimating the visibility of macroblock level distortion. Our approach is based on the simple observation that detailed textures tend to mask the visually perceived distortion level. It should be noted that the perceived impairment is also high if a textured block is replaced by a smooth block, or a smooth block is replaced by a textured block. Therefore, we will define spatial intensity S as a minimum spatial activity of the original and the replaced blocks:

$$S = \min(S_{\mathrm{REF}}, S_{\mathrm{TAR}}), \tag{1}$$

where S_{REF} is the spatial activity in the original reference block, and S_{TAR} is the respective spatial activity in the impaired target block. There are several ways to approximate the spatial activity level of a block of pixels. In ITU-R Recommendation P.910 [51], Sobel filter is first applied to the frame to reveal edges, and then standard deviation is computed and used as a spatial activity index SI. In our study, we have adopted similar definition for S, but at the block level:

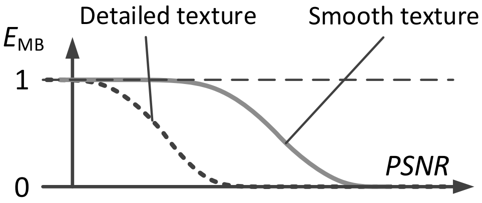


Fig. 3. Mapping of PSNR values into macroblock-level error visibility index (E_{MB}) depends on the spatial texture intensity S of the affected block.

$$S = \min\Big(\mathrm{std}\big[\mathrm{Sobel}\big(B_{2..13,2..13}\big)\big], \mathrm{std}\Big[\mathrm{Sobel}\Big(\hat{B}_{2..13,2..13}\Big)\Big]\Big), \tag{2}$$

where std() denotes standard deviation, Sobel() denotes Sobel filter operation as defined in [51], and $B_{i,j}$ is the pixel intensity (monochrome component) of reference block B at position (i, j), and $\hat{B}_{i,j}$ is the respective intensity in the impaired target block $\hat{B}$. Note that in our study, we use pixel intensities normalized to the range 0..1. We have also omitted two pixels that are closer than three pixels from the edges, in order to avoid edges at block border appearing due to temporal error concealment to impact the results; this is why the pixel indices for the 16×16 pixel blocks run 2..13 instead of 0..15.

In the following phase, $PSNR$ is computed for each macroblock, using the standard equation:

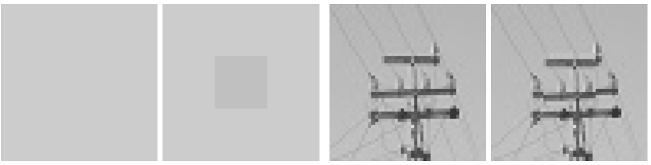
$$PSNR = 10 \cdot \log_{10}\left(1/\mathrm{mean}\left[\left(B_{0..15,0..15} - \hat{B}_{0..15,0..15}\right)^2\right]\right) \tag{3}$$

Finally, we can apply a sigmoid function to convert $PSNR$ and S into a macroblock-level error visibility index (E_{MB}):

$$E_{MB} = 1 - 1/(1 + \exp[\alpha \cdot S + \beta \cdot PSNR]) \tag{4}$$

The fixed parameters α and β will be defined empirically so that the subjective impairment matches with E_{MB}. Figure 3 illustrates the relationship between $PSNR$ and E_{MB}: in terms of PSNR, the visibility threshold depends on the intensity of the spatial texture, measured by T from Eq. (1). Values between 0 and 1 represent different levels of visibility: impairments of $E_{\mathrm{MB}} < 0.1$ are usually only visible in still images when inspected carefully, whereas impairments with $E_{\mathrm{MB}} > 0.9$ are highly visible even if only one isolated macroblock is impaired.

To find the optimal parameters α and β, we have used EPFL-PoliMi database [26]. The database consists of two datasets with test video sequences of two different resolutions: CIF (352×288 pixels) and 4CIF (704×576 pixels). We computed $PSNR$ and S for each macroblock in each of the 4CIF resolution test sequences in the database (CIF resolution dataset was not included, because CIF resolution is very low compared to our test video sequences of Full HD resolution). Then, we searched for the parameter values α and β that will minimize the linear correlation between the average E_{MB} and the respective MOS values for each sequence (since E_{MB} is a distortion score and MOS is a quality score,



a) Distortion on smooth surface MSE=0.0025; E_{MB} =0.174

b) Distortion on textured surface MSE=0.0252; E_{MB} =0.035

Fig. 4. Macroblock level visibility of distortions demonstrated on a smooth surface (a) and textured surface (b). Non-impaired version is on the left, and impaired version is on the right.

the ideal correlation between them would be −1). As a result, we obtained the optimal values $\alpha = -37$ and $\beta = -0.06$. Figure 4 shows an example of MSE and E_{MB} results in two different blocks with errors: as the example shows, MSE highly overestimates the visibility of a misaligned block on a detailed texture, whereas E_{MB} is much better in line with the subjectively perceived error visibility.

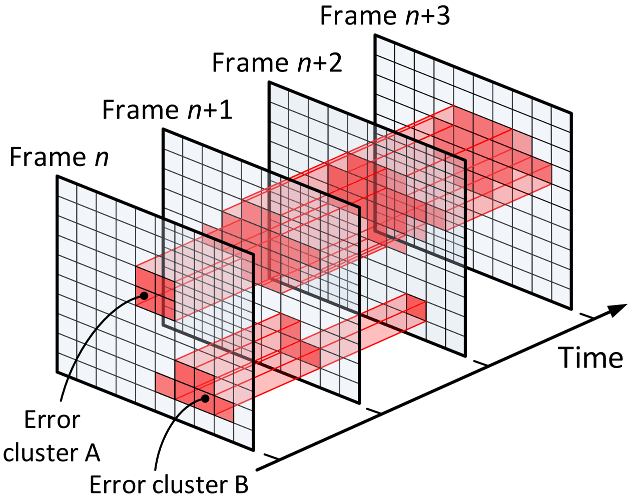


Fig. 5. Error clusters illustrated. Error cluster A is spreading along time, due to error propagation; error cluster B is vanishing, as prediction residuals gradually replace the distorted pixels.

C. Error Clusters

In the prior art, an individual packet loss artifact is usually defined as an impairment caused by a single packet loss. For example, in the work by Reibman and Poole [10], test sequences were created so that within each specific time interval, there was only one packet loss appearing. This is a reasonable approach, when the perceptual impact of an individual packet loss event is studied. However, in practical video sequences, packet losses are rarely isolated events. Very commonly, two or more packet losses have a visual impact in spatially and temporally overlapping regions. In this case, it is not always possible to visually distinguish the impact of separate packet loss events as separate artifacts. This is why we have taken a fundamentally different approach for defining a packet loss artifact.

We define an *error cluster* as an area in the spatiotemporal space with a high density of visual impairments caused by packet losses. The definition of error cluster is agnostic to actual packet loss events: an error cluster may appear as a result of one packet loss, or several packet losses. It is also possible that one packet loss causes two or more separate error clusters. For example, one packet loss can affect two visually complex objects on a smooth surface; in this case, the objects can suffer from visible impairments, but the smooth surface between the objects is efficiently concealed by an error concealment algorithm, and therefore the visible artifacts occur in two spatially separated locations.

Figure 5 shows an example with two separate error clusters. In this example, red blocks indicate visually impaired macroblocks. When visual impairments appear in adjacent frames in near spatial position, the impairments are supposed to belong to the same error cluster. Therefore, error clusters can be considered as three-dimensional bodies floating in the spatiotemporal space. The example in the Fig. 5 shows two error clusters appearing simultaneously, error cluster A spanning over four frames and error cluster B over three frames. Due to temporal prediction and error propagation, it is common for error clusters to change shape and position along time. In some cases, two (or more) error clusters can merge into one along time, or split into two (or more) spatially separate clusters.

To combine erroneous macroblocks into error clusters, certain rules need to be defined followed to avoid generating a large number of severely fragmented clusters. The algorithm is similar to that described in our earlier study [12], except that we use E_{MB} as a macroblock level indicator for error visibility, instead of MSE. First, E_{MB} is computed for all the macroblocks in the video sequence. Then, the algorithm goes through all the macroblocks and tests the following conditions:

1) If the average E_{MB} value in the surrounding window of 7×3 macroblocks is higher than θ_1, all macroblocks in that surrounding window are marked as erroneous.

2) If the average E_{MB} value in the surrounding window of 5×3 macroblocks is higher than θ_2, all macroblocks in that surrounding window are marked as erroneous.

3) If the average E_{MB} value in the surrounding window of 3×3 macroblocks is higher than θ_3, all macroblocks in that surrounding window are marked as erroneous.

4) If the E_{MB} value is higher than θ_4, all macroblocks in the 3×3 surrounding window are marked as erroneous.

Asymmetric windows are used, because typical coding pattern of macroblocks follows horizontal rather than vertical order, and horizontal edges are more prominent than vertical edges in many practical video contents (such as landscape with a horizon). If the macroblock is located close to the border of the frame, the surrounding window sizes are adjusted so that it fits to the valid area. Since the subjective error visibility is related to human perception, there is no analytical method to determine the threshold values. If the thresholds are too high, there will be visible impairments that are not included in any error cluster. On the other hand, if the thresholds are too low, there will be a lot of small fragmented error clusters with hardly visible impairments.

In our implementation, we have used values $\theta_1 = \theta_2 = \theta_3 = 0.1$, and $\theta_4 = 0.25$, determined by trial and error.

Algorithm 1 Find if MB is Classified as Impaired

```
for n ≔ 0:L-1          // Go through all frames
  for i ≔ 0:M-1        // X-axis
    for j ≔ 0:N-1      // Y-axis

      // Indices for 3x3, 5x3 and 7x3 windows
      idx_y = max(0,j-1):min(N-1,j+1)
      idx_x_1 = max(0,i-1):min(M-1,i+1)
      idx_x_2 = max(0,i-2):min(M-1,i+2)
      idx_x_3 = max(0,i-3):min(M-1,i+3)

      // Test whether e_mb[i,j,n] belongs to an
      // error cluster
      if mean(e_mb[idx_x_3,idx_y,n])>theta_1
        is_err[idx_x_3,idx_y,n] ≔ true
      elseif mean(e_mb[idx_x_2,idx_y,n])>theta_2
        is_err[idx_x_2,idx_y,n] ≔ true
      elseif mean(e_mb[idx_x_1,idx_y,n])>theta_3
        is_err[idx_x_1,idx_y,n] ≔ true
      elseif e_mb[i,j,n] > theta_4
        is_err[idx_x_1,idx_y,n] ≔ true
      else
        is_impaired [i,j,n] ≔ false
      endif

    endfor
  endfor
endfor
```

The implementation of the algorithm in pseudocode is shown below (algorithm 1). Distortion index in the macroblock in position (i, j) of frame n is `e_mb[i,j,n]`, `M` and `N` denote the spatial dimensions of the frame (in macroblocks) and `L` denotes the length of the sequence (in frames). Binary flag `is_err[i,j,n]` will indicate if the macroblock in position (i, j) is classified as impaired or not.

After the first pass, the algorithm will go through the macroblocks again to assign error cluster identifiers to the macroblocks that are marked as impaired. When two impaired macroblocks are located next to each other, they are assigned the same identifier, i.e., they belong to the same error cluster. Also, if two error clusters in consecutive frames are fully or partially overlapping in the spatial dimension, the same identifier will be used for them, and this is how the error clusters can span over several frames (as seen in Fig. 5). As a special case, two or more error clusters can also merge along time: in this case, the merged error cluster is considered as continuation of the largest of the error cluster in the previous frame(s). In some cases, error clusters can also split into smaller parts. In this case, they are considered to be part of the original error cluster (i.e., the same identifier is assigned to them).

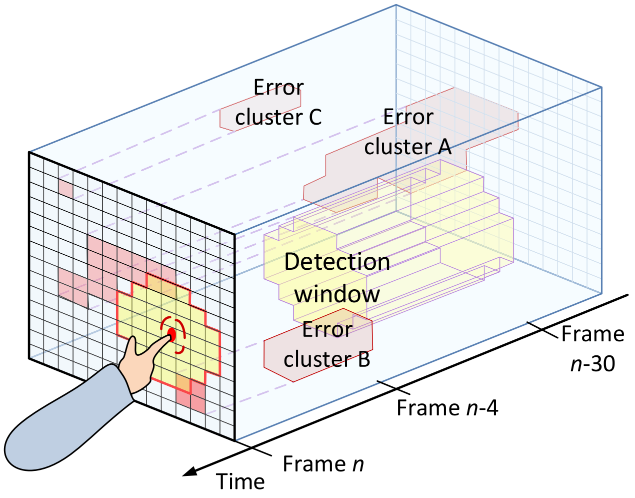


Fig. 6. Example of a detection window in the spatiotemporal space, overlapping with error clusters A and B.

D. Assigning Taps to Error Clusters

The aim of the subjective experiment using the touchscreen was to study the visibility of error clusters. For this purpose, each tap should be assigned to the most likely error cluster the user has observed. To relate each tap to a spatiotemporal region, we have defined a *detection window:* spatially, it spans symmetrically around the tapped macroblock, including 37 macroblocks in total. Temporally, detection window spans from frame $n - 30$ to $n - 4$, where n is the tapped frame. With 25 frames per second, it is equivalent to the time interval starting 1200 milliseconds before the tap and ending 160 milliseconds before the tap. We assume that this interval will cover typical reaction times. In the relevant related studies, average reaction times of 400-500 milliseconds have been observed [29], [52], and a vast majority of test subjects reacts faster than in one second. Note that for different frame rates, the interval has to be adjusted accordingly.

Since the test video sequence contains a large number of error clusters, a detection window may overlap with several error clusters. This is illustrated in Fig. 6, where detection window is denoted in yellow color and error clusters in red color. In this example, error clusters A and B overlap with the detection window in the spatiotemporal space. Error cluster C overlaps with the detection window only temporally, not spatially, and it is therefore not considered as detected. It is a matter of definition, if both error clusters A and B should be considered detected, or only the one that has the most significant overlap with the detection window. In our initial study [12], we used spatiotemporal weighting to emphasize positions close to the spatial midpoint and temporally most likely reaction times within the detection window. We also made a distinction between “main detection” (the error cluster with the largest overlap with the detection window) and “side detections” (the error cluster(s) with smaller overlap(s) with the detection window).

In practice, inaccuracies in cluster formation and the physical act of tapping the screen seem to play a bigger role in causing misdetections than confusion between main and side detections. This is why we have simplified the technique for cluster detection by omitting the spatiotemporal weighting and the concept of main and side detections in this paper. Therefore, we have only counted the main detections (the error clusters with the largest overlap with the detection window without weighting). Each error cluster is classified as detected or non-detected in a binary fashion, even if the cluster is detected several times by separate taps. The most visible error clusters typically last for at least 200-300 milliseconds, which leaves enough time for test subjects to detect multiple error clusters appearing in different spatial position within the same detection window.

IV. Characterization of Packet Loss Artifacts

Several different measures can be defined to characterize error clusters. We can expect that the spatiotemporal size and the average distortion would be highly related with the subjective visibility of an error cluster, but other factors may have an influence, too. In this section, we analyze different characteristics of error clusters and how they are related to error visibility, and we also propose general error visibility models combining different features of error clusters. We have used MATLAB (R2017a) for analyzing the video sequences, and Python for training and validating the learning-based regression models.

A. Characteristics of Error Clusters

The most obvious attribute of an error cluster is its size in temporal and spatial dimensions, i.e., the number of macroblocks the cluster contains. The size can be divided into temporal length (number of frames) and average spatial size (total number of macroblocks divided by the number of frames). We have also computed the relative size of the error cluster, in respect with the temporally overlapping parts of other error clusters. The relative size is the number of macroblocks in the error cluster divided by the number of macroblocks in all the error clusters in the same frames.

Another important attribute is the intensity of distortion. In our preliminary analysis in [12], we used PSNR as a distortion measure, but in this paper, we will also compute different cluster level error distortion indices by combining E_{MB} values of the cluster by with different pooling schemes (maximum, mean, median, and average of the 10%, 25% and 50% of the macroblocks with the largest E_{MB}). We assume that content adaptive percentage pooling would give the most accurate results, but to keep the model reasonably simple, we have only tested those listed fixed percentages. We would not expect any essential improvement by using adaptive percentage pooling.

In addition to distortion measures, we have also computed the spatial and temporal activity indices (SI and TI) for each error cluster, following the definitions of SI and TI from [51]. Large SI indicates high spatial activity, i.e., detailed textures or a lot of edges etc. Large TI indicates high temporal activity, i.e., intensive motion. Our initial results suggested that there is a negative correlation between SI and visibility of the error cluster. This is expected, because anomalies are easier to observe on smooth surface than on detailed textures. There is also a positive correlation between TI and visibility of the error cluster. We assume that the combination of motion and defect attracts attention easier than static distortions. To capture the joint impact of spatial and temporal activity, we have also formulated a combined spatiotemporal activity index as $TI/(SI + 10^{-4})$, where the constant term 10^{-4} is used for the denominator to avoid division by zero when SI is zero.

To analyze how different characteristics influence the visibility of the error cluster, we have computed Pearson Linear Correlation Coefficient (PLCC) and Spearman Rank Order Correlation Coefficient (SROCC) between different attributes and subjective error cluster visibility (i.e., the proportion of test subjects who have detected the cluster). In addition, we have computed the p-values from PLCC ($n = 6487$) to test the statistical significance level of the observed correlation. The results are shown in Tables I-III, grouped in three main categories: attributes related to the size of the cluster are listed in Table I, distortion measures in Table II, and content related factors in Table III.

TABLE I
Correlation Between Error Cluster Size Measures and Subjective Error Visibility

Abbr.	Attribute explanation	PLCC	SROCC	p-value
TS	Temporal size (frames)	0.515	0.356	$<10^{-4}$
SS	Spatiotemporal size (MBs)	0.433	0.359	$<10^{-4}$
SS/*TS*	Spatial size (MBs)	0.271	0.298	$<10^{-4}$
RS	Relative size	0.526	0.365	$<10^{-4}$

TABLE II
Correlation Between Error Cluster Distortion Measures and Subjective Error Visibility

Abbr.	Attribute explanation	PLCC	SROCC	p-value
PSNR	Peak Signal-to-Noise Ratio	-0.123	-0.107	$<10^{-4}$
1/*PSNR*	Reciprocal for PSNR	0.125	0.107	$<10^{-4}$
max E_{MB}	Maximum E_{MB}	0.367	0.305	$<10^{-4}$
mean E_{MB}	Mean E_{MB}	0.101	0.113	$<10^{-4}$
median E_{MB}	Median E_{MB}	0.030	0.053	0.016
E_{MB}10%	Mean of 10% highest E_{MB}	0.244	0.215	$<10^{-4}$
E_{MB}25%	Mean of 25% highest E_{MB}	0.218	0.215	$<10^{-4}$
E_{MB}50%	Mean of 50% highest E_{MB}	0.175	0.165	$<10^{-4}$

TABLE III
Correlation Between Error Cluster Content Measures and Subjective Error Visibility

Abbr.	Attribute explanation	PLCC	SROCC	p-value
SI	Spatial activity	-0.106	-0.133	$<10^{-4}$
TI	Temporal activity	0.077	0.093	$<10^{-4}$
TI /(*SI*+10^{-4})	Spatiotemporal index	0.156	0.159	$<10^{-4}$

As expected, the results in Table I show that there is a relatively significant positive correlation between error cluster size and subjective visibility. Temporal length (number of frames) and spatiotemporal size (number of macroblocks) show roughly similar correlation. We have also computed the relative size for each error cluster, defined as the absolute spatiotemporal size of the cluster divided by the number of all the impaired macroblocks located in the respective frames. The relative size is also positively correlated with the subjective visibility.

The results in Table II show that there is a clear correlation between different distortion measures and the subjective error visibility. Apart from median E_{MB}, there is correlation for all the measures at a statistically significant level ($p < 10^{-4}$). The results indicate that E_{MB} is more accurate measure of subjectively observed distortion than PSNR; however, appropriate pooling has to be applied to generate the overall distortion index from the macroblock level indices. Maximum E_{MB}

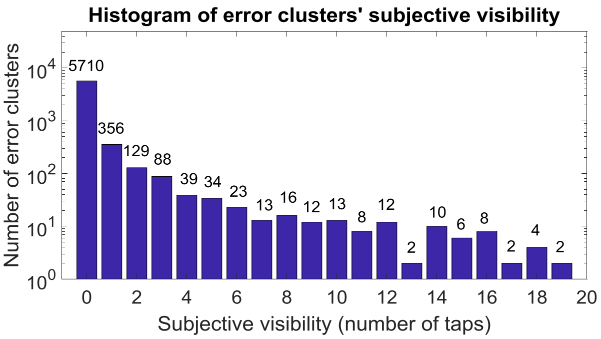


Fig. 7. Histogram for the error clusters classified by the number of detections (note the logarithmic scale for the Y-axis).

shows the highest correlation, followed by mean of 10% of the highest E_{MB} values in the cluster.

Table III shows that spatial and temporal activity levels of in the distorted regions also play a role for the error visibility. Reciprocal of spatial activity index SI shows higher correlation than temporal activity index TI, but both correlations are at the statistically significant level. By combining SI and TI into a spatiotemporal index $TI/(SI + 10^{-4})$, a stronger correlation can be observed.

We can conclude that the subjective visibility of an error cluster is contributed by all the studied factors: cluster size (both absolute and relative), strength of the distortion, and spatiotemporal activity of the original content in the region covered by the error cluster. The p-values show that the statistical significance level is high for the correlations computed for all the tested attributes, except median E_{MB}. However, none of the studied attributes alone can be used to predict the subjective error visibility level accurately.

It should be noted that correlation analysis between different attributes and the subjective visibility is challenging, due to the prevalence of very small error clusters. The small error clusters, comprising only few macroblocks, are most likely not noticeable; however, some of them may have been accidentally detected by erroneous or late taps. This may be the case particularly when a small error cluster is spatiotemporally located in the vicinity of a larger cluster. In Fig. 7, we have plotted the histogram of the clusters with different subjective error visibility levels. Most of the error clusters have not been detected by any subject, i.e., their subjective visibility is zero. On the other hand, none of the error clusters have been tapped by all the 20 test subjects; the highest subjective error visibility is 19. The histogram shows roughly a logarithmic distribution of subjective visibility levels: only 22 error clusters have been detected by 15 test persons or more (indicating high error visibility), 143 error clusters have been detected by five to 14 test persons (indicating intermediate error visibility), and remaining 6322 error clusters have been detected by four or less test persons (indicating low error visibility).

B. Analytical Model for Error Visibility

We have tried different approaches to develop an error cluster visibility metric, based on the individual attributes listed in Tables I-III. The most promising approach to compute cluster level error visibility index E_{CL} is to take a logarithm of the product of the n most relevant attributes $x_{1..n}$:

$$E_{\text{CL}} = \lg\left(\prod_{i=1..n} x_i\right) \tag{5}$$

TABLE IV
CORRELATION BETWEEN E_{CL} AND SUBJECTIVE ERROR VISIBILITY, USING DIFFERENT COMBINATIONS OF ATTRIBUTES TO COMPUTE E_{CL}

x_1	x_2	x_3	x_4	PLCC	SROCC	MSE
TS	E_{MB}10%	*TI*	*RS*	0.700	0.385	0.0035
TS	E_{MB}10%	1/(*SI*+10^{-4})	*RS*	0.732	0.402	0.0032
TS	max E_{MB}	*TI*/(*SI*+10^{-4})	*RS*	0.741	0.432	0.0031
SS	E_{MB}10%	*TI*	*RS*	0.667	0.379	0.0038
SS	E_{MB}10%	1/(*SI*+10^{-4})	*RS*	0.731	0.400	0.0032
SS	E_{MB}10%	*TI*/(*SI*+10^{-4})	*RS*	0.739	0.400	0.0031
SS	(E_{MB}10%)^2	*TI*/(*SI*+10^{-4})	*RS*	**0.748**	**0.407**	**0.0030**

We have tried the proposed formula with different combinations of attributes. Our observations show that in general, the strongest correlations are achieved by combining four attributes representing the following categories: absolute size, relative size, distortion intensity and spatiotemporal activity. In general, the attributes showing strongest correlations in Tables I-III typically also work best when used in Eq. (5) to compute E_{CL}, but there are exceptions to the rule: in particular, 10% percentage pooling for E_{MB} tends to work better than maximum E_{MB}.

Since the subjective error visibility is limited to the range from 0 to 1, we have used a piecewise function $f(E_{\text{CL}})$ to fit to the data, with parameters a and b ($a < b$) defining the lower and upper limits. In addition, parameter c is used as an exponent to better match with the shape of the values. The formulation of function $f(E_{\text{CL}})$ is as follows:

$$f(OEVI) = \begin{cases} 0, & \text{if } OEVI \leq a \\ (OEVI - a)/(b - a), & \text{if } a < OEVI < b \\ 1, & \text{if } OEVI \geq b \end{cases} \tag{6}$$

In Table IV, we have listed the best results in terms of PLCC, SROCC and MSE, when different sets of attributes are applied to Eq. (5), and minimum least squares (MLS) regression is used to fit the piecewise function defined in Eq. (6) to the data. Since there are a large number of different distortion measures as listed in Table II, we have included only some of the best performing distortion metrics with different combination in Table IV.

In terms of SROCC, the best result was obtained by using combination of *SS* for the spatiotemporal size, maximum E_{MB} value for distortion, *TI*/(*SI*) for a content-based coefficient, and the relative size as a fourth attribute. If PLCC or MSE is used as a performance criterion, a better result was achieved by using a combination of SS, mean of the 10% highest E_{MB} values squared, spatiotemporal index $TI/(SI + 10^{-4})$, and relative size *RS*. Due to the limited number of error subjective visibility levels in the data, we can consider PLCC and MSE as more appropriate performance metric than SROCC; therefore, the last line in Table IV is taken as the best performing combination of attributes.

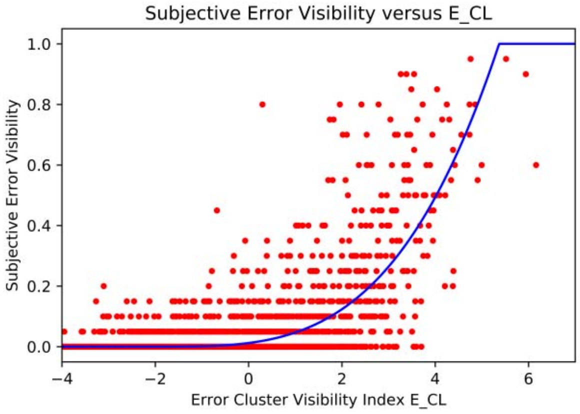


Fig. 8. Subjective error visibility as a function of the cluster level error visibility, computed from Eq. (5). Red dots denote the individual error clusters, and blue line shows the nonlinear fit to the data, using Eq. (6).

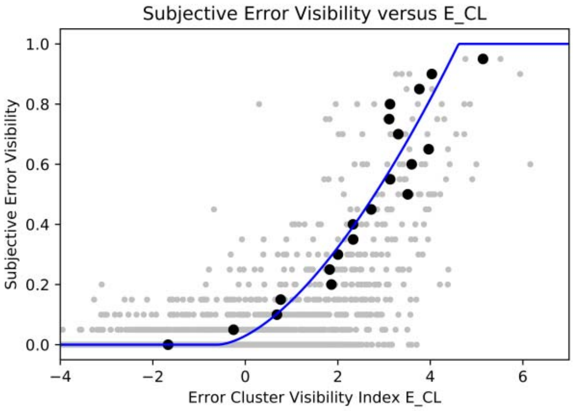


Fig. 9. Subjective error visibility as a function of the average cluster level error visibility computed for each level. Grey dots denote the individual error clusters, black dots denote the average E_{CL} for each level, and blue line shows the nonlinear fit to the averages, using Eq. (6).

The results in Table IV show that the prediction accuracy for subjective visibility will be significantly improved by computing E_{CL} with different combinations of attributes, compared to the individual attributes listed in Tables I-III. The correlation coefficients are still not very strong. However, as shown in the histogram in Fig. 7, the distribution of the subjective error visibility levels is very imbalanced. Due to the nature of the subjective experiment, the data is also rather noisy. This is why it is expected that there are relatively large amount of error clusters that are detected by a mistake, or a tap is erroneously assigned to a less visible error cluster that is spatiotemporally close to a more visible error cluster.

In Fig. 8, we have plotted the subjective error visibility levels as a function of E_{CL}, computed using the best performing combination of attributes (written in bold in Table IV). The blue line shows the piecewise nonlinear fit to the data. As we can see, the curve fits reasonably well to the data points with low error visibility level, but it predicts the data points with a high error visibility level less accurately. Since the vast majority of the error clusters have very low subjective visibility, minimum least squares regression overemphasizes the clusters with low error visibility.

In order to give equal weight for different error visibility levels, we have computed the average E_{CL} values for each subjective visibility level. Then, we applied the piecewise function to the average values instead of the raw data points. Note that there are no error clusters detected by every test subject, and this is why the subjective visibility level 1 is omitted; therefore, there are only 19 data points representing different subjective visibility levels from zero to 0.95. The average E_{CL} values for each subjective visibility level are plotted in Fig. 9.

The average E_{CL} values can be fitted with high accuracy by piecewise linear regression (PLCC = 0.942, SROCC = 0.926, MSE = 0.0094). When the model is applied to the original test data, correlation coefficients PLCC = 0.721, SROCC = 0.468 and MSE = 0.0072 are obtained. In terms of SROCC, it is better than the result when regression is applied to the original data, and in terms of PLCC, it is only slightly worse. However, MSE is noticeably worse, since the subjective visibility of the low visibility level error clusters are predicted less accurately, and those clusters significantly outnumber the clusters with high error visibility. On the other hand, the data concerning the low error visibility clusters is rather noisy, and in spite of worse overall MSE, the model with higher discriminatory power regarding highly visible error clusters may be preferred in practical applications.

C. Machine Learning Model for Error Visibility

Regression models based on machine learning are highly popular in practical applications of computer vision, including visual quality assessment. This is why we have also tried different machine learning techniques to predict subjective visibility from a set of attributes selected from those listed in Tables I–III. In the literature, several different machine learning techniques have been applied for image and video quality assessment, including Support Vector Machine (SVM), Convolutional Neural Network (CNN) and Random Forest (RF) regression [53]–[56].

Our dataset is particularly challenging for machine learning, because it is produced in a specific experiment and therefore rather limited. This excludes data intensive techniques, such as CNN and deep learning. The output values (subjective error visibility levels) are also very unbalanced. An ideal model would be able to discriminate low and high visibility error clusters accurately, but this would require that the data points with different visibility levels were relatively evenly represented in the training and validation datasets. A common practice in machine learning is to use a random split of 80:20 into training and validation data. In our case, this would mean that the validation data would contain very few data points representing the high error visibility clusters. Even if 50:50 random split is used, the high error visibility clusters may be very unevenly distributed, since there are so few of them.

To guarantee a fair balance of data points in training and validation data, we have split each of the 21 error visibility levels into training and validation data separately, using 50:50 split for each level. In this way, we can make sure that each error visibility level is represented in training and

TABLE V
PERFORMANCE OF DIFFERENT REGRESSION METHODS COMPARED

	Average			Standard deviation		
Regressor	PLCC	SROCC	MSE	PLCC	SROCC	MSE
SVM	0.614	0.353	7.7e-3	0.021	0.011	8.9e-4
MLP	0.650	0.364	4.1e-3	0.028	0.008	2.2e-4
RF	0.727	0.373	3.3e-3	**0.013**	0.012	**1.4e-4**
GB	0.749	0.391	3.1e-3	0.017	0.009	1.9e-4
Bagging	**0.761**	0.395	**2.9e-3**	0.018	**0.008**	1.9e-4
Analytical	0.748	**0.407**	3.0e-3	-	-	-

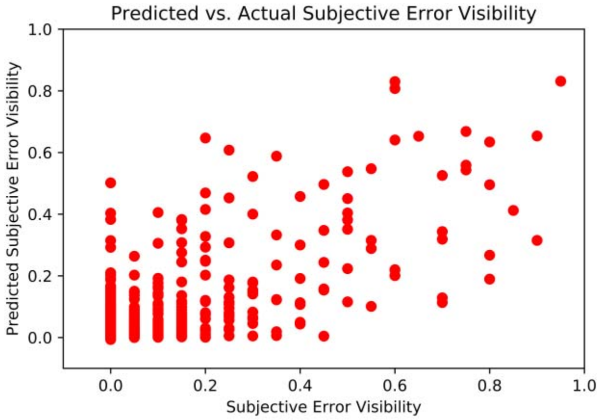


Fig. 10. Accuracy of bagging regression for predicting the subjective visibility levels illustrated. Red dots denote individual error clusters.

validation sets, even though there are only few data points for some levels (e.g., two data points for levels 13, 17 and 19). The input features are pre-processed by rescaling them to interval 0.1.

We have tried different random splits and different regression models available in scikit-learn toolbox for Python to predict subjective error visibility level from six features, namely TS, SS, RS, SI, TI and $E_{\mathrm{MB}}10\%$ (see Tables I-III). These are the same features we have used for the analytical model described in Section IV-B. Since our dataset is unbalanced, the random split into training and test sets has a relatively high impact on the results. This is why we have tried 25 different splits, and reported the average results (PLCC, SROCC and MSE), as well as standard deviations, in Table V. Four different regressors were used: Multi-Layer Perceptron (MLP), Support Vector Machine (SVM) with Gaussian kernel, Random Forest (RF), and Gradient Boosting (GB). Default kernels and parameters were used for each technique. We have also included results using Bagging Regression with GB as a base estimator. As a comparison, the results obtained from the analytical model are also reported.

As the results show, GB regressor performs the best, and its performance can be further boosted by using bagging. Standard deviations indicate that the best performing techniques also give relatively constant results across different splits to training and validation data. An example of the prediction results using bagging regressor is illustrated in Fig. 10.

D. Discussion

As expected, spatial and temporal extent of the error cluster is correlated with its subjective visibility. Also the relative size of the error cluster (in respect with the other error clusters that are present in the overlapping temporal window) shows a clear positive correlation with the subjective error visibility. However, the importance of the size should not be emphasized too much: even a dummy test subject tapping the screen randomly would be expected to detect some of the large error clusters, since random detection windows are more likely to overlap with large error clusters than small error clusters. In some cases, even large error clusters may go unnoticed, if the intensity of the distortion is low.

In our prior study, we used conventional PSNR to measure the intensity of the distortion. In this paper, we have proposed a macroblock level error visibility index E_{MB} that predicts the perceived distortion more accurately than PSNR. We have tried different pooling methods to combine macroblock level E_{MB} values into a cluster level distortion measure E_{CL}; correlation analysis shows that the highest correlation with subjective error visibility can be achieved from the mean of the highest 10% of the E_{MB} values. We can conclude that E_{MB} is substantially more accurate metric for estimating the visibility of packet loss artifacts than PSNR.

The video content also impacts the visibility of errors. When motion is present, temporal error propagation often causes artifacts that look unnatural and change shape along time. Temporally intensive artifacts are therefore more likely to draw attention than static artifacts in temporal static regions of video. This is why temporal intensity level is positively correlated with error visibility. Spatial activity is also related to error visibility. Errors on very detailed textures are less noticeable than errors on more smooth regions, since the fine details often "masks" the distortion. This effect is considered already in E_{MB}. Nevertheless, including spatial activity index into the overall error visibility estimate improves the correlation with subjective visibility. By integrating any sophisticated saliency model (see examples in [57]), we could expect some performance improvement. However, it would be questionable if the improvement would be sufficient to justify the increased complexity.

The best performing regression techniques based on machine learning work only slightly better than the simple analytical model we have derived. However, we assume that the regression results could be improved further by using a more extensive dataset and by selecting the kernel functions and parameters more carefully. However, given the inaccuracies in the subjective experiment, as well as the lack of a precise definition for an error cluster, the results can be considered satisfactory already, and drastic improvement would not be expected.

The proposed features for characterizing packet loss artifacts are relatively computationally very simple. On the other hand, the proposed algorithm for combining erroneous macroblocks into error clusters requires multiple passes, and it is therefore computationally more complex. Our test implementation (in MATLAB) is not capable for real-time processing of video sequences, but we assume that a properly optimized implementation in a lower level compiled programming language (such as C or C++) could be used for continuous monitoring of received video signal in real time.

V. Conclusion

In this paper, we have studied packet loss distortion in video sequences by analyzing individual, spatiotemporally limited error clusters. We have extended the analysis of the subjective study published in [12] by deriving new numerical features characterizing the error clusters and improving the method for determining the error clusters. We have also proposed two new models for estimating the subjective visibility of error clusters from those features: an analytical model, and a learning-based model. We have shown that there are several different factors influencing the visibility of packet loss artifacts. Spatiotemporal extent of the impaired region, as well as the intensity of distortion are important factors, but also the content, defined in terms of spatial and temporal intensity, plays a role on the subjective visibility. In the related studies, the focus is usually on the video quality on the sequence level, not on the visibility of different isolated artifacts individually. This is why we believe that our study provides valuable insights for researchers developing methods for minimizing the impact of packet losses in video streaming and broadcasting applications. We also expect that the proposed FR method for detecting error clusters can be useful as a benchmark to NR methods for detecting packet loss artifacts at macroblock level.

References

[1] "Methodology for the subjective assessment of the quality of television pictures," Int. Telecommun. Union, Geneva, Switzerland, ITU-Recommendation BT.500-13, 2012.

[2] O. B. Maua, H. C. Yehia, and L. de Errico, "A concise review of the quality of experience assessment for video streaming," *Comput. Commun.*, vol. 57, no. 2, pp. 1–12, Feb. 2015, doi: 10.1016/j.comcom.2014.11.005.

[3] L. Krasula, P. Le Callet, K. Fliegel, and M. Klíma, "Quality assessment of sharpened images: Challenges, methodology, and objective metrics," *IEEE Trans. Image Process.*, vol. 26, no. 3, pp. 1496–1508, Mar. 2017, doi: 10.1109/TIP.2017.2651374.

[4] J.-S. Lee, "On designing paired comparison experiments for subjective multimedia quality assessment," *IEEE Trans. Multimedia*, vol. 16, no. 2, pp. 564–571, Feb. 2014, doi: 10.1109/TMM.2013.2292590.

[5] J. Korhonen, C. Mantel, and S. Forchhammer, "Subjective comparison of brightness preservation methods for local backlight dimming displays," in *Proc. SPIE/IS T EI*, San Francisco, CA, USA, 2015, Art. no. 939504.

[6] J. Tompkin, M. H. Kim, K. I. Kim, J. Kautz, and C. Theobalt, "Preference and artifact analysis for video transitions of places," *ACM Trans. Appl. Percept.*, vol. 10, no. 3, p. 13, Aug. 2013, doi: 10.1145/2501601.

[7] T. Alpert and J.-P. Evain, "Subjective quality evaluation—The SSCQE and DSCQE methodologies," in *Proc. EBU Tech. Rev.*, 1997, pp. 12–20. [Online]. Available: https://tech.ebu.ch/docs/techreview/trev_271-evain.pdf

[8] A. Borowiak, U. Reiter, and U. P. Svensson, "Quality evaluation of long duration audiovisual content," in *Proc. CCNC*, Las Vegas, NV, USA, 2012, pp. 337–341.

[9] S. Jumisko-Pyykkö, V. M. Kumar, and J. Korhonen, "Unacceptability of instantaneous errors in mobile television: From annoying audio to video," in *Proc. MobileHCI*, Helsinki, Finland, 2006, pp. 1–8.

[10] A. R. Reibman and D. Poole, "Predicting packet-loss visibility using scene characteristics," in *Proc. PV*, Lausanne, Switzerland, 2007, pp. 308–317.

[11] S. Argyropoulos, A. Raake, M.-N. Garcia, and P. List, "No-reference video quality assessment for SD and HD H.264/AVC sequences based on continuous estimates of packet loss visibility," in *Proc. QoMEX*, Mechelen, Belgium, 2011, pp. 31–36.

[12] J. Korhonen and C. Mantel, "Assessing visibility of individual transmission errors in networked video," in *Proc. IS T EI/HVEI*, San Francisco, CA, USA, 2016, p. 8.

[13] G. Valenzise, S. Magni, M. Tagliasacchi, and S. Tubaro, "No-reference pixel video quality monitoring of channel-induced distortion," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 22, no. 4, pp. 605–618, Apr. 2012, doi: 10.1109/TCSVT.2011.2171211.

[14] J. Koloda, J. Østergaard, S. H. Jensen, V. Sánchez, and A. M. Peinado, "Sequential error concealment for video/images by sparse linear prediction," *IEEE Trans. Multimedia*, vol. 15, no. 4, pp. 957–969, Jun. 2013, doi: 10.1109/TMM.2013.2238524.

[15] J. Liu, G. Zhai, X. Yang, B. Yang, and L. Chen, "Spatial error concealment with an adaptive linear predictor," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 25, no. 3, pp. 353–366, Mar. 2015, doi: 10.1109/TCSVT.2014.2359145.

[16] M. Ebdelli, O. Le Meur, and C. Guillemot, "Video inpainting with short-term windows: Application to object removal and error concealment," *IEEE Trans. Image Process.*, vol. 24, no. 10, pp. 3034–3047, Oct. 2015, doi: 10.1109/TIP.2015.2437193.

[17] B. Wang, Q. Peng, J. Chen, and P. Gao, "A low-complexity error concealment algorithm for video transmission based on non-local means denoising," in *Proc. VCIP*, Chengdu, China, Nov. 2016, pp. 1–4.

[18] Y.-K. Wang, M. M. Hannuksela, V. Varsa, A. Hourunranta, and A. Gabbouj, "The error concealment feature in the H.26L test model," in *Proc. ICIP*, Rochester, NY, USA, 2002, pp. 729–732.

[19] P. Lambert, W. De Neve, Y. Dhondt, and R. Van de Walle, "Flexible macroblock ordering in H.264/AVC," *J. Vis. Commun. Image Represent.*, vol. 17, no. 2, pp. 358–375, Apr. 2006, doi: 10.1016/j.jvcir.2005.05.008.

[20] S. Wenger, "H.264/AVC over IP," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 13, no. 7, pp. 645–656, Jul. 2003, doi: 10.1109/TCSVT.2003.814966.

[21] J. Chakareski, J. G. Apostolopoulos, S. Wee, W. Tan, and B. Girod, "Rate-distortion hint tracks for adaptive video streaming," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 15, no. 10, pp. 1257–1269, Oct. 2005, doi: 10.1109/TCSVT.2005.854227.

[22] Z. Chen and D. Wu, "Prediction of transmission distortion for wireless video communication: Analysis," *IEEE Trans. Image Process.*, vol. 21, no. 3, pp. 1123–1137, Mar. 2012, doi: 10.1109/TIP.2011.2168411.

[23] S. Tang and P. R. Alface, "Impact of random and burst packet losses on H.264 scalable video coding," *IEEE Trans. Multimedia*, vol. 16, no. 8, pp. 2256–2269, Dec. 2014, doi: 10.1109/TMM.2014.2348947.

[24] S. Winkler and R. Campos, "Video quality evaluation for Internet streaming applications," in *Proc. SPIE (HVEI)*, Santa Clara, CA, USA, 2003, pp. 104–115.

[25] T. Liu, Y. Wang, J. M. Boyce, Z. Wu, and H. Yang, "Subjective quality evaluation of decoded video in the presence of packet losses," in *Proc. ICASSP*, Honolulu, HI, USA, 2007, pp. 1125–1128.

[26] F. De Simone *et al.*, "Subjective quality assessment of H.264/AVC video streaming with packet losses," *EURASIP J. Image Video Process.*, vol. 2011, Dec. 2011, Art. no. 190431, doi: 10.1155/2011/190431.

[27] J. Nightingale, Q. Wang, C. Grecos, and S. Goma, "The impact of network impairment on quality of experience (QoE) in H.265/HEVC video streaming," *IEEE Trans. Consum. Electron.*, vol. 60, no. 2, pp. 242–250, May 2014, doi: 10.1109/TCE.2014.6852000.

[28] K. Seshadrinathan, R. Soundararajan, A. C. Bovik, and L. K. Cormack, "Study of subjective and objective quality assessment of video," *IEEE Trans. Image Process.*, vol. 19, no. 6, pp. 1427–1441, Jun. 2010, doi: 10.1109/TIP.2010.2042111.

[29] H. Wang *et al.*, "VideoSet: A large-scale compressed video quality dataset based on JND measurement," *J. Vis. Commun. Image Represent.*, vol. 46, pp. 292–302, Jul. 2017, doi: 10.1016/j.jvcir.2017.04.009.

[30] S. Kanumuri, P. C. Cosman, A. R. Reibman, and V. A. Vaishampayan, "Modeling packet-loss visibility in MPEG-2 video," *IEEE Trans. Multimedia*, vol. 8, no. 2, pp. 341–355, Apr. 2006, doi: 10.1109/TMM.2005.864343.

[31] S. Kanumuri, S. G. Subramanian, P. C. Cosman, and A. R. Reibman, "Packet-loss visibility in H.264 videos using a reduced reference method," in *Proc. ICIP*, Atlanta, GA, USA, 2006, pp. 2245–2248.

[32] Z. Wang, L. Lu, and A. C. Bovik, "Video quality assessment based on structural distortion measurement," *Signal Process. Image Commun.*, vol. 19, no. 2, pp. 121–132, Feb. 2004, doi: 10.1016/S0923-5965(03)00076-6.

[33] M. H. Pinson and S. Wolf, "A new standardized method for objectively measuring video quality," *IEEE Trans. Broadcast.*, vol. 50, no. 3, pp. 312–322, Sep. 2004.

[34] K. Seshadrinathan and 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, doi: 10.1109/TIP.2009.2034992.

[35] 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. 587–599, Apr. 2010, doi: 10.1109/TCSVT.2010.2041829.

[36] F. Tommasi, V. De Luca, and C. Melle, "Packet losses and objective video quality metrics in H.264 video streaming," *J. Vis. Commun. Represent.*, vol. 27, no. 2, pp. 7–27, Feb. 2015, doi: 10.1016/j.jvcir.2014.12.003.

[37] U. Reiter, J. Korhonen, and J. You, "Comparing apples and oranges: Assessment of the relative video quality in the presence of different types of distortions," *EURASIP J. Image Video Process.*, vol. 2011, p. 8, Sep. 2011, doi: 10.1186/1687-5281-2011-8.

[38] I. Sedano, M. Kihl, K. Brunnström, and A. Aurelius, "Evaluation of video quality metrics on transmission distortions in H.264 coded video," in *Proc. BMSB*, Nuremberg, Germany, 2011, pp. 1–5.

[39] M. Shahid, A. Rossholm, B. Lövström, and H.-J. Zepernick, "No-reference image and video quality assessment: A classification and review of recent approaches," *EURASIP J. Image Video Process.*, vol. 2014, p. 40, Dec. 2014, doi: 10.1186/1687-5281-2014-40.

[40] M. T. Vega, D. C. Mocanu, S. Stavrou, and A. Liotta, "Predictive no-reference assessment of video quality," *Signal Process. Image Commun.*, vol. 52, pp. 20–32, Mar. 2017, doi: 10.1016/j.image.2016.12.001.

[41] K. Zhu, C. Li, V. Asari, and D. Saupe, "No-reference video quality assessment based on artifact measurement and statistical analysis," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 25, no. 4, pp. 533–546, Apr. 2015, doi: 10.1109/TCSVT.2014.2363737.

[42] A. Barri and A. Dooms, "Data-driven modules for objective visual quality assessment focusing on benchmarking and SLAs," *IEEE J. Sel. Topics Signal Process.*, vol. 11, no. 1, pp. 196–205, Feb. 2017.

[43] N. Montard and P. Bretillon, "Objective quality monitoring issues in digital broadcasting networks," *IEEE Trans. Broadcast.*, vol. 51, no. 3, pp. 269–275, Sep. 2005, doi: 10.1109/TBC.2005.851700.

[44] H. Rui, C. Li, and S. Qiu, "Evaluation of packet loss impairment on streaming video," *J. Zhejiang Univ. Sci. B*, vol. 7, pp. 131–136, Jan. 2006, doi: 10.1631/jzus.2006.AS0131.

[45] E. P. Ong *et al.*, "Video quality monitoring of streamed videos," in *Proc. ICASSP*, Taipei, Taiwan, 2009, pp. 1153–1156.

[46] N. Teslic, V. Zlokolica, V. Pekovic, T. Teckan, and M. Temerinac, "Packet-loss error detection system for DTV and set-top box functional testing," *IEEE Trans. Consum. Electron.*, vol. 53, no. 3, pp. 1312–1319, Aug. 2010, doi: 10.1109/TCE.2010.5606264.

[47] D. Shabtay, N. Raviv, and Y. Moshe, "Video packet loss concealment detection based on image content," in *Proc. EUSIPCO*, Lausanne, Switzerland, 2008, pp. 1–8.

[48] I. Glavota, M. Vranješ, M. Herceg, and R. Grbić, "Pixel-based statistical analysis of packet loss artifact features," in *Proc. ZINC*, Novi Sad, Serbia, 2016, pp. 16–19.

[49] M. H. Pinson, "The consumer digital video library [best of the Web]," *IEEE Signal Process. Mag.*, vol. 30, no. 4, pp. 172–174, Jul. 2013, doi: 10.1109/MSP.2013.2258265.

[50] *H.264/AVC Reference Software*. Jul. 11 2014. [Online]. Available: iphome.hhi.de/suehring/tml/

[51] "Subjective video quality assessment methods for multimedia applications," Int. Telecommun. Union, Geneva, Switzerland, ITU-Recommendation P.910, 1999.

[52] M. W. G. Dye, C. S. Green, and D. Bavelier, "The development of attention skills in action video game players," *Neuropsychologia*, vol. 47, nos. 8–9, pp. 1780–1789, Jul. 2009, doi: 10.1016/j.neuropsychologia.2009.02.002.

[53] J. Søgaard, S. Forchhammer, and J. Korhonen, "No-reference video quality assessment using codec analysis," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 25, no. 10, pp. 1637–1650, Oct. 2015, doi: 10.1109/TCSVT.2015.2397207.

[54] S.-C. Pei and L.-H. Chen, "Image quality assessment using human visual DOG model fused with random forest," *IEEE Trans. Image Process.*, vol. 24, no. 11, pp. 3282–3292, Nov. 2015, doi: 10.1109/TIP.2015.2440172.

[55] Y. B. Youssef, A. Mellouk, M. Afif, and S. Tabbane, "Video quality assessment based on statistical selection approach for QoE factors dependency," in *Proc. GLOBECOM*, Washington, DC, USA, 2016, pp. 1–6, doi: 10.1109/GLOCOM.2016.7842375.

[56] J. Li, J. Yan, D. Deng, W. Shi, and S. Deng, "No-reference image quality assessment based on hybrid model," *Signal Image Video Process.*, vol. 11, no. 6, pp. 985–992, Sep. 2017, doi: 10.1007/s11760-016-1048-5.

[57] X. Wang, L. Ma, S. Kwong, and Y. Zhou, "Quaternion representation based visual saliency for stereoscopic image quality assessment," *Signal Process.*, vol. 145, pp. 202–213, Apr. 2018, doi: 10.1016/j.sigpro.2017.12.002.

Jari Korhonen (M'05) received the M.Sc. (Eng.) degree in information engineering from the University of Oulu, Oulu, Finland, in 2001 and the Ph.D. degree in telecommunications from the Tampere University of Technology, Tampere, Finland, in 2006. He is currently with the Institute of Future Media Technology, Shenzhen University, China, where he has been a Research Assistant Professor since 2017.

From 2001 to 2006, he was a Research Engineer with Nokia Research Center, Tampere. He was with the École Polytechnique Fédérale de Lausanne, Switzerland, in 2007 and the Norwegian University of Science and Technology, Trondheim, Norway, from 2008 to 2010. From 2010 to 2017, he was with the Technical University of Denmark. His research interests include both telecommunications and signal processing aspects in multimedia communications, including visual quality assessment.