DOI: 10.2352/ISSN.2470-1173.2016.16HVEI-106

Assessing Visibility of Individual Transmission Errors in Networked Video

Jari Korhonen, and Claire Mantel; Technical University of Denmark; Lyngby, Denmark

Abstract

Traditionally, subjective video quality is assessed by user experiments involving quality ratings, pairwise comparisons, or rank ordering, based on the overall impression of quality. Less attention has been paid on assessing the visibility of individual defects. However, many practical applications could benefit from information about subjective visibility of individual packet losses; for example, computational resources could be directed more efficiently to unequal error protection and concealment by focusing in the visually most disturbing artifacts. In this paper, we present a novel subjective methodology for packet loss artifact detection by tapping a touchscreen where a defect is observed. To validate the proposed methodology, the results of a pilot study are presented and analyzed. According to the results, the proposed method can be used to derive qualitatively and statistically meaningful data on the subjective visibility of individual packet loss artifacts.

Introduction

Visual quality assessment is important for many practical applications, such as quality-based classification, optimization of coding parameters, error protection and post-processing, in particular error concealment. In an ideal case, visual quality would be measured objectively, i.e. by an algorithm that analyzes features of the content and generates automatically a quality index that is well in line with the subjectively experienced quality. In practice, objective quality measurement is not a trivial task, and subjective quality assessment studies are required to find the *ground truth* information about the subjective quality.

Conventionally, test subjects in subjective quality assessment studies rate the shown content using a given rating scale [1]. When a number of subjects have given their individual scores, Mean Opinion Score (MOS) can be computed as an average of the individual scores. MOS is then considered as a subjective quality indicator of the content. Different rating scales can be used: a five-point scale ranging from one ("very poor") to five ("excellent") is very common, but there are also different scales, including differential scales where the test content is compared against a reference content and rated using a relative scale (for example, from "much worse than reference" to "much better than reference") [2]. Recently, pairwise comparison methods have gained popularity [3], and rank ordering methods are also occasionally used [4]. In general, there is no consensus about the universally best subjective test method in the scientific community; the preferred test method depends on the purpose of the test, content type, distortion type and many other factors.

Unfortunately, a single score given for a video clip gives little information on how the different defects contribute to the overall score. In networked video content, visual quality is often affected by both compression and transmission artifacts. In adaptive streaming, compression rate may vary as a result of changing available bandwidth, which will cause temporal variations in the visual quality. Transmission errors (such as packet losses) are often concentrated into clusters that are scattered unevenly in both spatial and temporal dimensions.

There are several reasons why the temporal dynamics of video artifacts are important for practical applications. Firstly, the quality level during the last few seconds of a test sequence may be overly emphasized in subjective ratings, since the short-term memory has a limited range [5]. However, objective metrics should treat all parts of the sequence equally, since real-life viewing of television or streaming video often contains breaks and disruptions, and therefore it is not directly comparable to a subjective quality assessment task. Secondly, fluctuating quality is typically judged subjectively more disturbing than constant quality, even if the average quality was somewhat lower [6]. Thirdly, spatiotemporal differentiation of distortions would be useful for certain applications, such as perceptually driven unequal error protection or prioritized error concealment.

To capture the dynamic nature of visual quality better than the conventional rating schemes, methods for temporally continuous rating have been proposed. For example, in Single Stimulus Continuous Quality Evaluation (SSCQE) method, a slider is used to rate the content continuously as the subjective quality changes [7]. Jumisko-Pyykkö et. al. studied acceptability of mobile audiovisual content by employing a button that is pressed when the quality drops to an unacceptable level [8]. Borowiak et. al. proposed another method where the quality of a test sequence is gradually deteriorated, and the task for the test person is to adjust the quality back to its original level by turning a control knob [9]. Kanumuri et. al. assessed visibility of packet losses by using a method where the space bar is pressed when a packet loss artifact is observed [10-11].

In spite of its benefits, temporally continuous quality assessment is used relatively rarely in the related research. One reason may be the challenging analysis of temporal quality data. Continuous quality assessment is also a more demanding task than giving one score for a sequence, and the continuous data is therefore often more noisy than the individual scores. Lastly, temporally continuous assessment does not encompass the spatial dimension and the interplay between the spatial and temporal dynamics of visual artifacts that are particularly essential for packet loss artifacts, typically appearing in spatially and temporally restricted regions. This is why we propose a method for indicating transmission artifacts in a video sequence by tapping a touchscreen where the artifact appears. The proposed method allows us to gather more precise information about error visibility in different regions of the image than a continuous quality score. In addition, the relative visibility of errors appearing approximately at the same time can be assessed. This is the main difference between our method and the method used in [8,10-11].

DOI: 10.2352/ISSN.2470-1173.2016.16HVEI-106

The rest of this paper is organized as follows. First, we explain the proposed subjective test methodology, how the content was generated, and how the practical test was arranged. Then, we present our approach for analyzing the results: first, we define the error clusters caused by packet losses, and then, we relate taps on the touchscreen to error clusters. Finally, we summarize the results from the practical study, and the concluding remarks are given.

Subjective Test

In order to assess the visibility of individual error clusters subjectively, we have designed a test protocol where a test video is displayed on a touchscreen, and a test person is instructed to tap the screen in the position where a packet loss error appears. Since test persons are not necessarily familiar with the concept of packet loss error, we prepared also a short introduction video with some examples of packet loss artifacts. Test persons could ask questions after watching the introduction video, before starting the test. Before the actual test video, a short training sequence was displayed, to allow test persons to familiarize with the test protocol. To avoid reflections of light on the display, the test was run in a room with the lights switched off and windows blinded by curtains.

Content generation

In conventional quality rating experiments, several short test clips are typically shown and rated during a subjective experiment. Since our experiment is by nature continuous, we decided to use 19 test sequences produced by Swedish Television (SVT) as source content. The sequences are freely available for research use in Consumer Video Digital Library (CVDL) [12]. When played sequentially without breaks between, the sequences form a continuous storyline with several different scenes with various levels of details and types of motion, such as panning and zooming, well representing a typical television viewing experience. Original sequences are in YUV 4:4:4 format with Full HD resolution (1920x1080 pixels) and a frame rate of 50 frames per second. However, to facilitate processing, we converted the sequences to YUV 4:2:0 format and reduced the frame rate to 25 frames per second. The total length of the video, and therefore that of the test, is about six minutes.

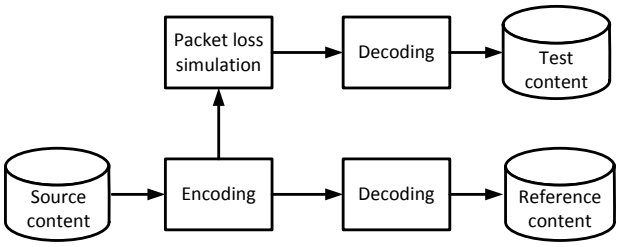

Figure 1. The workflow for generating the test sequences.

To create the test sequences, we first encoded the sequences using H.264/AVC reference codec (version 12.4) [13]. We set the group of pictures (GOP) length to 25, with the frame structure IBBPBBP… for temporal prediction. Flexible macroblock (MB) ordering (FMO) with the standard chessboard pattern was enabled, since it is commonly used also in real-life digital television content. Each Network Adaptation Layer Unit (NALU) was configured to include a fixed number of MBs (200), to make it easier to generate packet losses at certain spatial locations. To ensure roughly constant quality without excessive compression artifacts, quantization parameter (QP) was set to 24, instead of using rate control.

Our intention was to create test sequences with realistic packet loss artifacts, where the impact of several individual packet losses may be intertwined, due to the complex spatial and temporal prediction structures between NALUs. In the related studies on this topic, only one packet loss is typically allowed in each timeslot of a few seconds. In that case, it is possible to relate each press of a button with a certain packet loss [14]. Since we will also collect the spatial location of the observed artifacts, we can omit this limitation, and let multiple packet losses to occur in each GOP.

Packet losses were simulated by a Matlab script removing NALUs randomly from the encoded sequences. The process was partially supervised, in order to keep the amount of visible packet loss artifacts at a meaningful level. For example, if authors' informal inspection of the damaged content revealed that the losses in certain regions caused excessive impact, the loss rate in those regions was reduced accordingly. The average packet loss rate was approximately 1.5%, and since there is 42 NALUs per frame, each GOP is impacted by approximately 15 NALU losses in average. The procedure for creating the test sequences and the error-free reference sequences is illustrated in Figure 1.

Practical study

In order to validate the test methodology and gather preliminary experiences, we arranged a small scale pilot study employing the proposed methodology. Twenty persons participated in the study, 7 females and 13 males, aged between 20 and 33. All except one of the subjects self-reported vision (for both acuity and color) that is normal or normal with glasses. None of the test persons indicated substantial experience on image processing. The test subjects were informed about the purpose of the test (i.e. to gather statistical information about the visibility of packet loss artifacts). They were also told that there would be a large number of artifacts and they were not expected to detect them all, so it was best to concentrate on the most obvious ones.

The test was run by using dedicated software written specifically for this purpose in C++ using Qt Creator platform, and compiled for Windows 7 operating system. The software displayed the test clips sequentially and recorded information about taps (coordinates and frame number) in a results file. Starting and ending times of the test were also recorded. There was no audio track or any kind of sound effects in use. As hardware we used a Dell's panel PC with a 21.5 inch touchscreen of Full HD resolution, equipped with a flash solid state drive as a mass memory, to allow sufficiently fast reading of video data in raw YUV format. The height of the display is 29 cm, and since the touchscreen had to be located close enough to the test person to be tapped conveniently, the distance to the display was approximately 1.5-2 times the height.

Analysis of Results

The workflow for analyzing the data is outlined in Figure 2. The test sequence is compared against the reference sequence to find the error clusters, as described below. At this phase, data characterizing each error cluster can also be computed. When the results from subjective tests are available, different analysis techniques can be used to correlate the taps with error clusters. Since the methodology is novel, there are no established

DOI: 10.2352/ISSN.2470-1173.2016.16HVEI-106

techniques for this purpose, and we have therefore developed our own approach.

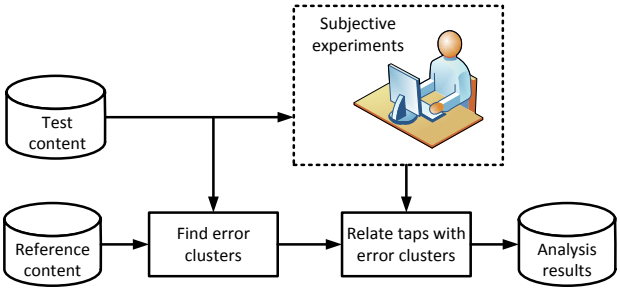

Figure 2. The workflow for analyzing the results.

Finding error clusters

Appearance of packet loss defects varies extensively, based on the video codec, coding parameters (such as temporal prediction length and slice structure), error concealment method (advanced techniques can conceal impact of lost MBs better than simple ones) and even content type (for example, complex textures can "mask" the impact of errors). Due to temporal prediction, errors in intra frames (I-frames) are often propagated to the predicted frames (P- and B-frames). Spatial prediction techniques may spread packet loss artifacts beyond the borders of the slices that were actually lost. If FMO is used, a checkerboard pattern with damaged and error-free MBs alternating will appear. The adopted error concealment strategy will also influence the appearance of the artifacts resulting from packet loss.

The shape and extent of an artifact caused by an individual packet loss is difficult to predict. When several packets are lost, covering an overlapping spatiotemporal area, it is also difficult to link the contribution of each lost packet to the overall quality degradation in the affected region. This is why we do not use a packet loss as a basic unit of error. Instead, we have adopted a concept of *error cluster* to describe a spatiotemporally restricted area, where distorted pixels appear. The distortions within each error cluster can be contributed by one or several packet losses.

The first challenge is to create a definition for an error cluster. As many common video coding standards, such as MPEG-2 and AVC/H.264, use 16x16 pixel MBs as a basic unit for coding and temporal prediction, our analysis is performed on per-MB basis; each MB is classified as non-erroneous (i.e. it does not belong to any error cluster) or erroneous, when it forms an error cluster with the other erroneous macroblocks located spatially or temporally next to it. Sometimes there may be two or several error clusters located close to one another, and in this case it is not always trivial to decide whether a group of erroneous MBs should be classified as a one large cluster or several smaller clusters. After attempting several different approaches, we have decided to use the following procedure for finding the error clusters:

1) Primary and secondary visibility thresholds for an error is defined in terms of *mean squared error* (MSE) per MB, denoted as t_1 and t_2, respectively ($t_1 > t_2$).
2) If a MB, or any other MB within the surrounding window of 3x3 MBs (to avoid "holes" inside clusters), has MSE larger than t_1, then this MB is classified as erroneous.
3) If the average MSE of all the neighboring MBs within any of the windows of size 3x3, 5x3 or 7x3 MBs is larger than t_2, then the MB is classified as erroneous. The reason for using those shapes is that typical slices span further in horizontal than vertical direction.
4) Any two (or more) erroneous MBs located vertically or horizontally next to each other belong to the same error cluster.
5) If the considered MB is erroneous, and also the MB located in the same position in the previous frame was erroneous, they belong to the same error cluster.
6) Sometimes two error clusters grow together along time. After merging, the joint error cluster is assigned to the cluster that was the largest in the previous frame.

We have implemented an algorithm in Matlab that follows the procedure described above. The algorithm uses as input the reference sequence (decoded video without packet losses) and the test sequence (decoded video corrupted by packet losses). In our implementation, we have used the detection thresholds t_1=0.001 and t_2=0.0001 computed on luma pixels (values normalized to interval 0-1), respective to *peak signal-to-noise ratio* (PSNR) values 30 dB and 40 dB. As an output, the algorithm produces a three-dimensional matrix with dimensions (w,h,f), where w is the width of a frame in MBs, h is the height in MBs, and f is the number of frames in the sequence. In the output matrix, each element represents a MB that is marked as unaffected (0), or belonging to an error cluster identified by a sequence number (n).

An example of detected error clusters in three consecutive frames is shown in Figure 3. The damaged frames are shown on top, a distortion map of damaged MBs is shown in the middle, and the error clusters found by the algorithm described above are shown on the bottom row. The distortion map is created so that badly distorted MBs (PSNR less than 20 dB) are white and non-distorted MBs (PSNR higher than 50 dB) are shown in black. PSNR values between 20 and 50 dB are represented by different gray levels, ranging from full white to full black.

In some cases, error has been detected, but it is not visible to a human observer. For example, there is a severely distorted area detected in the lower part of the third frame in Fig. 3, even though it is efficiently concealed by copying the respective macroblocks from the previous frame. PSNR for that region is low, since the fine texture of the grass is misplaced in respect with the undamaged reference frame. However, as the figure shows, it is difficult or impossible for a human eye to see any defect (at least in a static image). In some other locations, for example around the right foot of the player on the right, misplacement of a macroblock causes a clearly visible artifact.

Correlating taps with error clusters

In the prior art [10-11], each detected error (i.e. the press of a button) can be easily related to a packet loss, since each packet loss event has been separated from other packet loss events by a time interval longer than a typical reaction time. However, it is a more challenging task to relate individual taps to error clusters in our study, since the error clusters have very different and uneven shapes and distributions of distorted pixels. In addition, the human reaction time and spatial accuracy of taps varies between different test persons.

DOI: 10.2352/ISSN.2470-1173.2016.16HVEI-106

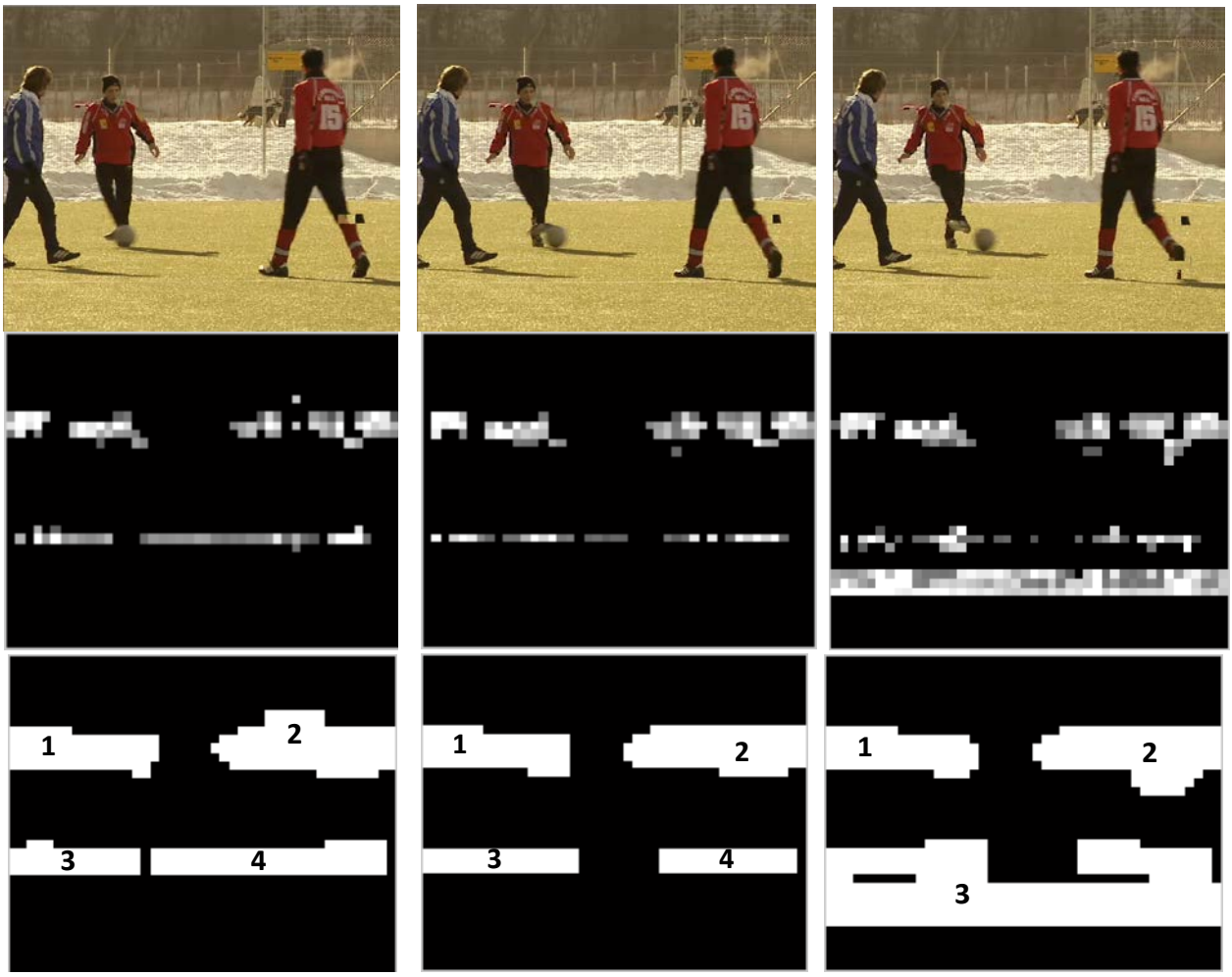

Figure 3. Three consecutive video frames with packet loss artifacts (top), heatmap of erroneous MBs (middle), and the detected error clusters (below).

We have chosen a straightforward technique where a *detection window* is defined for each tap. If there is any MB belonging to a certain error cluster inside the detection window, that error cluster is marked as detected. In our analysis, we have used a detection window that includes 37 MBs around the tapped MB (see the shape in Figure 4) and all the frames from *z*-25 to *z*-5, where frame *z* is when the tap was recorded. Frames *z*-4…*z* are not included in the detection window, because we assume that the reaction time is always more than 200 ms. As the physical size of a MB on the used test screen is about 4.2x4.2 millimeters, an error cluster will be marked as detected if the screen is tapped closer than 12.6 millimeters from any MB of that error cluster, and no longer than one second after the cluster has disappeared. Figure 4 shows a graphical illustration of the detection window.

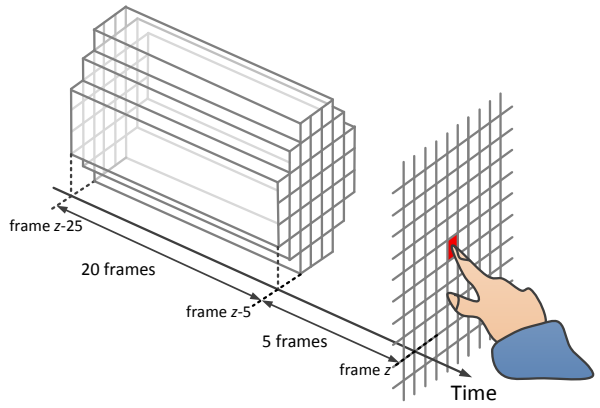

Figure 4. Detection window for correlating taps with error clusters.

The binary decision between detected and non-detected error clusters carries some problems. First, there will be unavoidably some small error clusters that will be marked as detected only because they are located spatiotemporally close to a large cluster. Second, some error clusters span over a long sequence of frames (even several seconds), and they may be tapped several times, indicating higher visibility, which is omitted in the binary decision. More comprehensive analysis may be performed by using a weighted detection index D_n for error cluster n, defined as:

©2016 Society for Imaging Science and Technology
DOI: 10.2352/ISSN.2470-1173.2016.16HVEI-106

$$D_n = \sum_{(i,j,k)\in\Omega} w_{i,j,k} \cdot m_n(i,j,k), \quad (1)$$

where Ω is the detection window, i and j are the spatial coordinates and k is the frame index, and $w_{i,j,k}$ are the weights based on the proximity to the tapped MB, $m_n(i,j,k)$=1 if MB at position (i,j,k) belongs to the error cluster n, and $m_n(i,j,k)$=0 if it does not. Basically, higher D_n suggests higher probability that error cluster n has been detected. This can be used in different ways: a detection threshold could be defined, or if there are several error clusters detected by the same tap, only the one with highest D_n would be marked as detected.

The weights $w_{i,j,k}$ are defined so that the highest values are assigned to the positions where the test person was most likely intending to tap. In our analysis, we have first defined spatial weights as:

$$S = \begin{bmatrix} 0 & 0 & 0.1 & 0.1 & 0.1 & 0 & 0 \\ 0 & 0.2 & 0.3 & 0.4 & 0.3 & 0.2 & 0 \\ 0.1 & 0.3 & 0.6 & 0.8 & 0.6 & 0.3 & 0.1 \\ 0.1 & 0.4 & 0.8 & 1.0 & 0.8 & 0.4 & 0.1 \\ 0.1 & 0.3 & 0.6 & 0.8 & 0.6 & 0.3 & 0.1 \\ 0 & 0.2 & 0.3 & 0.4 & 0.3 & 0.2 & 0 \\ 0 & 0 & 0.1 & 0.1 & 0.1 & 0 & 0 \end{bmatrix} \quad (2)$$

In S, the middle position has the highest weight (1), as it represents the MB that was tapped. The surrounding MBs have lower weights, following roughly the shape of a Gaussian function. To obtain the final weights, we have defines temporal weighting function $T(k)$:

$$T(k) = \begin{cases} 0, \text{ if } k > z\text{-}5 \\ \left(\frac{k+26-z}{22}\right)\left(1-\frac{k+26-z}{22}\right)^{1/2}, \text{if } z\text{-}5 \geq k \geq z-25 \\ 0, \text{if } k < z\text{-}25 \end{cases} \quad (3)$$

We have formulated the temporal weighting function so that it approximates the shape of the probability distribution for reaction times. Since we do not know the typical human reaction times in this particular task, we have assumed that the reaction times are roughly similar to those reported in the context of video gaming [15], typically ranging from 400 to 500 ms. Roughly similar average reaction times have been observed in [14], but with a sharper distribution; we assume that in our study the distribution of reaction times is wider, since the task of locating the error and tapping the screen is more complex than just pressing a button. This is why we did not use directly the data from [14] or [15] for our weighting function. The weights derived from (3) are shown in Figure 5.

When the MB at position (x,y) in frame z is tapped, the weights $w_{i,j,k}$ can be obtained by combining the spatial weights from matrix S and the temporal weight as:

$$w_{i,j,k} = s_{x-i+3,y-i+3} \cdot T(k) \quad (4)$$

An alternative approach for correlating taps with errors would be to create a heatmap of objectively detected distortions and another heatmap for subjectively detected distortions (taps), and then study their correlation in a similar fashion as heatmaps derived by saliency models are compared against heatmaps generated from eye tracking data [16]. As an objective heatmap, the distortion map computed on a per-MB basis (as shown in Fig. 3) could be used. A subjective heatmap could be generated by using Eq. (4) to create distortion maps derived from each tap, and then combining distortion maps derived from different taps into an overall heatmap. However, since each tap on a touchscreen only approximates the location of the distorted regions, not their shapes and extents, we assume that it is more meaningful to relate taps with error clusters instead of a heatmap. Analysis based on heatmaps may be considered in the future research.

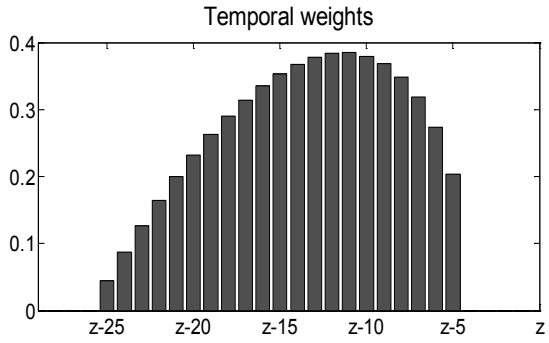

Figure 5. Temporal weights for frames z-25 to z, computed from (3).

Practical test results

The test content, composed of 19 test sequences, contains 7896 error clusters in 9061 frames. The vast majority of the error clusters are very small; only 3456 clusters contain more than nine MBs. On the other hand, the largest cluster contains more than 180,000 MBs and spans temporally over 110 frames (more than four seconds). As a comparison, one frame contains 8,160 MBs. In total, 9.4% of all MBs are classified to belong to an error cluster. It should be noted that different algorithms for cluster detection could produce very different results. However, since we do not have *a priori* knowledge about the visibility of clusters, there is no meaningful method for comparing different cluster detection algorithms. The subjective test results may be used for improving the cluster detection in an iterative manner, but at this stage of our study, this is considered as part of future research.

Large differences in the number of taps and detected error clusters between test persons were observed. Apparently, some test subjects took a more "competitive" approach in the experiment, attempting to tap as many errors as possible, whereas some others were more focused on the story and tapped the most obvious defects only. Nevertheless, we assume that by averaging the results from different persons, meaningful indicators of subjective visibility of error clusters can be derived. Table I lists some key statistics of the experiment: a) the number of taps, b) the number of detected error clusters, c) the number of detected error clusters when only the cluster with the highest weight is counted per tap, and d) missed taps (i.e. taps that did not hit any error cluster). Minimum, average, median and maximum values from different test persons are given. Note that some large clusters may have been tapped several times, but each detected cluster is counted only once in the statistics.

DOI: 10.2352/ISSN.2470-1173.2016.16HVEI-106

Table I: Statistical summary for the number of taps, detected clusters (all), detected clusters (only the strongest counted per tap) and missed taps.

	a)	b)	c)	d)
minimum	82	56	33	4
median	165.5	190.0	123.5	17.5
average	211.1	198.0	124.3	23.8
maximum	654	360	199	73
standard dev.	127.3	74.6	40.3	17.6

As demonstrated in Figure 3, majority of the error clusters are not highly noticeable to human observers. In total, 1083 error clusters were detected by at least one test person, but many of those are small clusters located in the same detection window with a larger cluster; we can refer to them as "side detections". If only one cluster, the one with the highest weight per tap, is counted, 483 clusters were detected. Out of those, 175 were detected by one test person only, 94 by at least 10 test persons (i.e. detection probability is 50% or higher), and only one cluster was detected by all the twenty test persons. From these results, we can conclude that there are approximately 90 clearly visible error clusters that will be observed by the majority of people.

To analyze further the correlations between individual test persons, we computed Pearson correlation coefficients between the detections by each test person (zero is non-detected and one is detected), and the average number of detections by the others (i.e. the number of other test persons who detected the cluster). This allows us to compare the individual detection accuracy against the other persons' detection accuracy. We omitted the clusters not detected by any of the test persons, as well as side detections. The resulting coefficients ranged from 0.374 to 0.675 for nineteen persons, with average 0.545. For one test person, a significantly lower correlation coefficient 0.027 was found, indicating an outlier. Due to the nature of the experiment, the data was supposed to be noisy, but the reasonably strong positive correlation coefficients suggest that the test arrangement is meaningful. Stronger correlations can be observed, if weighted detection indices are used instead of the binary detection data.

We have computed some basic indicators to characterize the error clusters, namely the length (in frames), spatial size (average number of MBs per frame), and distortion (in PSNR). In addition, we have computed spatial activity index (SI) and temporal activity index (TI) for each error cluster. High value for SI indicates high level of spatial details, and high value for TI indicates intensive motion, respectively. Instead of using the standard definitions for SI and TI from [17], we have redefined SI and TI as follows:

$$SI_n = mean\left[Sobel(C_n)\right] \tag{5}$$

$$TI_n = mean\left[C_n - C_{n,prev}\right] \tag{6}$$

In (5) and (6), n denotes the index of the error cluster, C_n includes all the pixels in those MBs that are marked to belong to the error cluster n, $C_{n,prev}$ includes the respective pixels in the preceding frame, and *Sobel* denotes the standard Sobel filter. Only the monochrome component (Y) of the original non-distorted frames is used for computing the PSNR, SI and TI values for each error cluster.

Since large error clusters may draw attention from smaller error clusters, we have also defined an indicator called proportion of erroneous MBs (PEM). PEM is simply the number of erroneous MBs that belong to a certain error cluster divided by the total number of erroneous MBs appearing in the same frames. Therefore, PEM=1 indicates that there are no other error clusters appearing in the same time interval. Low value for PEM indicates that there are a lot of erroneous MBs belonging to other error clusters, competing for viewer's attention. This is why we can hypothesize that PEM is positively correlated with error cluster's likelihood to be detected.

The average values of the chosen indicators (temporal length, spatial size, PSNR, SI, TI and PEM) for the detected and undetected error clusters are shown in Table II. In order to avoid noise from side detections, only primarily detected clusters (the one with the highest weight in the detection window) by at least one test person are considered as detected. The respective median values are shown in Table III.

Table II: Average characteristics of detected and undetected error clusters.

	length (frames)	spatial size	distortion (PSNR)	SI	TI	PEM
detected	31.81	227.9	31.60	4.80	2.93	0.167
undetected	3.86	14.7	38.75	5.33	1.57	0.034
all	5.59	27.9	38.30	5.30	1.65	0.053

Table III: Median characteristics of detected and undetected error clusters.

	length (frames)	spatial size	distortion (PSNR)	SI	TI	PEM
detected	29	99.2	32.43	4.17	2.26	0.067
undetected	1	4.0	37.98	3.77	1.49	0.006
all	1	4.0	37.84	3.79	1.52	0.008

As the results show, the detected clusters tend to be significantly longer and larger than the undetected clusters. Also the average PSNR is lower for the detected clusters than undetected clusters. These results are well in line with the results expected by intuition. For SI and TI, the results are more difficult to interpret. The mean SI is larger for undetected than detected clusters, whereas the median SI is larger for detected clusters. In average, TI for detected clusters seems to be higher than for undetected clusters. We will discuss the possible explanations in the following Section.

Different subjective visibility indices can be derived from the subjective data. Most obviously, the mean of the weighted detection indices D_n will be directly related to the expected visibility level. In addition, the average number of detections (number of test persons who detected the cluster) could be considered as a visibility index. In Table IV, Pearson correlation coefficients (PCC) are computed to estimate the relationship between subjective visibility (average number of detections and sum of weighted detection indices; side detections are omitted) and error cluster characteristics (temporal length, spatial size, distortion, SI, TI and PEM). Only the error clusters detected by at least one person are taken into consideration in the computations. When all the error clusters are included, the observed correlations are slightly stronger; however, in respect with each other, they are comparable with the results shown in Table IV. We did not observe any significant differences in the results, when different inclusion criteria for data points were used.

DOI: 10.2352/ISSN.2470-1173.2016.16HVEI-106

Table IV: Correlations (PCC) between subjective visibility indicators and objective characteristics of error clusters.

	avg. detections	mean(D_n)
length in frames	0.420	0.445
spatial size in MBs	0.368	0.376
PSNR	-0.390	-0.399
SI	-0.008	0.021
TI	0.280	0.292
PEM	0.459	0.458

As the results show, there is a moderate positive correlation between both visibility indicators and the temporal and spatial size of error clusters, which follows intuition. Negative correlation is observed between subjective visibility and PSNR, which is expected as well, since higher PSNR indicates lower distortion. It should be noted that each MB of an error cluster that is located inside the detection window will contribute to D_n. Since large clusters usually occupy larger part of the detection window than small clusters, the size of the cluster has an impact on D_n. This explains why spatiotemporal size has stronger correlation with mean value of D_n than the percentage of detections among test persons. In any case, we can conclude from the results that both spatiotemporal size of the cluster and distortion observed within the cluster (PSNR) contribute to the overall visibility of the error cluster. This observation seems trivial, but it is a necessary step for validating the proposed methodology.

Discussion and future research

In the related literature, several subjective studies have been reported, where packet loss artifacts are indicated along the temporal dimension, e.g. by pressing a button [8,10-11,18]. This kind of method is useful for analyzing the visibility of individual packet losses, when they are separated in different timeslots. However, in real-life high resolution video sequences, errors often appear in different spatial positions and overlap in temporal dimension. In our method, we use a touchscreen to record also the spatial position of the detected error. This allows us to analyze the relative visibility of temporally overlapping errors, which is not possible with the conventional methods. On the other hand, a larger number of dimensions also has disadvantages: our method has a higher cognitive load, and the results are more difficult to analyze. This is why we must accept greater inaccuracies in the results, compared to error detection performed solely in the temporal dimension.

The results of this study show that the proposed methodology can be used to produce credible indicators for subjective visibility, showing strong correlation with objective features that are expected to be related with error visibility. When this information is available, we expect that we can develop more accurate objective metrics predicting the subjective visibility of individual error clusters. The subjective visibility data could also be highly useful for validating no-reference algorithms for detecting packet loss artifacts in decoded video sequences. Some methods for packet loss artifact detection have been proposed in the literature [19,20], but since there is no subjective visibility data available for individual error clusters, those methods have only been evaluated by using objectively measured distortion (such as MSE) as ground truth. As demonstrated in Fig. 3, objective distortion measures may not be well in line with subjectively experienced distortion. This is the case particularly when individual error clusters are evaluated instead of the overall quality.

Our results show that the likelihood to detect an error cluster is clearly correlated with the spatiotemporal size of the error cluster, as well as PSNR. In terms of PCC, the correlation between error visibility and the SI is weak, even though the results show differences in mean and median values of SI for detected and undetected error clusters. This observation leads us to assume that error clusters with very high or very low spatial activity are less likely to be detected than error clusters with intermediate level of spatial activity. We assume that complex textures (high SI) can "mask" the impact of distortion, whereas error concealment effectively recovers distortions on smooth surfaces (low SI), which could explain the higher visibility of distortions located in regions with intermediate spatial details.

Positive (although relatively weak) correlation is observed between TI and error visibility. Possible reason is that distortions in static regions (low TI) are easier to conceal by spatial replacement than distortions in regions with intensive motion (high TI). It is evident that SI and TI do influence the error visibility; however, their impact is not as straightforward as the impact of the spatiotemporal size and PSNR. This is why SI and TI are not as strongly correlated with error visibility. In the future research, we will analyze the influence of spatiotemporal activity on the error visibility more comprehensively.

It should be noted that in our correlation analysis, we have included only error clusters that have been detected by at least one person. This is because the vast majority of error clusters are undetected. When undetected clusters are also included, somewhat higher correlation coefficients can be obtained between error visibility and temporal length, spatial size and TI. In contrast, the correlation between error visibility and PSNR becomes slightly lower. The same is observed if we include the side detections that were omitted in the analysis above. The differences are relatively small and do not compromise our observations and conclusions in general. In the future, we will study different methods to combine the features we have analyzed into a model that can be used to predict the visibility of error clusters more accurately.

Conclusions

In this paper, we have proposed a subjective test methodology for assessing the visibility of individual error clusters caused by packet losses in networked video. In the subjective experiment, the test person taps a touchscreen in a position where a packet loss defect is observed. We have conducted a pilot study to verify the methodology, and we have also proposed techniques for analyzing the data. Comparisons between subjective indicators of visibility, such as the percentage of test persons who have detected an error cluster, and objective indicators of visibility, such as spatiotemporal size and measured distortion level of an error cluster, show clear correlations, which suggests that the methodology is valid. In addition, the correlation statistics show a reasonable level of coherence between test persons, which also suggests that the results can be considered meaningful. However, a more extensive subjective study is required to obtain statistically more accurate data on subjective visibility. In the future research, objective models will be developed to predict subjective visibility of error clusters, using the subjective data generated from the proposed testing method as a ground truth for calibrating and validating the objective models.

DOI: 10.2352/ISSN.2470-1173.2016.16HVEI-106

References

[1] ITU-R Recommendation BT.500-13, "Methodology for the Subjective Assessment of the Quality of Television Pictures," 2012.

[2] O. B. Maua, H. C. Yehia, and L. de Errico, "A Concise Review of the Quality of Experience Assessment for Video Streaming," Computer Communications, vol. 68, no. 2, pp. 1-12, 2015.

[3] R. K. Mantiuk, A. Tomaszewska, and R. Mantiuk, "Comparison of Four Subjective Methods for Image Quality Assessment," Computer Graphics Forum, vol. 31, no. 8, pp. 2478-91, 2012.

[4] J. Korhonen, C. Mantel, and S. Forchhammer, "Subjective Comparison of Brightness Preservation Methods for Local Backlight Dimming Displays," in IS&T/SPIE Electronic Imaging Conference, San Francisco, California, 2015.

[5] D. S. Hands, and S. E. Avons, "Recency and Duration Neglect in Subjective Assessment of Television Picture Quality," Applied Cognitive Psychology, vol. 15, no. 6, pp. 639-657, 2001.

[6] S. Thakolsri, W. Kellerer, and E. Steinbach, "QoE-based Cross-layer Optimization of Wireless Video with Unperceivable Temporal Video Quality Fluctuation," in IEEE International Conference on Communications, Kyoto, Japan, 2011.

[7] T. Alpert, and J.-P. Evain, "Subjective Quality Evaluation – The SSCQE and DSCQE Methodologies," EBU Technical Review, pp. 12-20, Spring 1997.

[8] S. Jumisko-Pyykkö, V. Kumar, and J. Korhonen, "Unacceptability of Instantaneous Errors in Mobile Television: From Annoying Audio to Video," in ACM Conference on Human-Computer Interaction with Mobile Devices and Services, Helsinki, Finland, 2006.

[9] A. Borowiak, U. Reiter, and U. P. Svensson, "Quality Evaluation of Long Duration Audiovisual Content," in IEEE Consumer Communications and Networking Conference, Las Vegas, Nevada, 2012.

[10] S. Kanumuri, P. C. Cosman, A. R. Reibman, "Modeling Packet-Loss Visibility in MPEG-2 Video," IEEE Transactions on Multimedia, vol. 8, no. 4, pp. 341-355, 2006.

[11] S. Kanumuri, S. G. Subramanian, P. C. Cosman, A. R. Reibman, "Packet-Loss Visibility in H.264 Videos using a Reduced Reference Method," in IEEE International Conference on Image Processing, Atlanta, GA, USA, 2006.

[12] Consumer Video Digital Library, <www.cdvl.org>

[13] H.264/AVC Reference Software, <iphome.hhi.de/suehring/tml/>

[14] M. W. G. Dye, C. S. Green, D. Bavelier, "The Development of Attention Skills in Action Video Game Players," Neuropsychologia, vol. 47, no. 8-9, pp. 1780-89, 2009.

[15] T.-L. Lin, S. Kanumuri, Y. Zhi, D. Poole, P. C. Cosman, and A. R. Reibman, "A Versatile Model for Packet Loss Visibility and its Application to Packet Prioritization," IEEE Transactions on Image Processing, vol. 19, no.3, pp. 722-735, 2010.

[16] T. Judd, F. Durand, and A. Torralba,"Fixations on Low-Resolution Images," Journal of Vision, vol. 11, no. 4, pp. 1-20, 2011.

[17] ITU-R Rec. P.910, "Subjective Video Quality Assessment Methods for Multimedia Applications," International Telecommunication Union, Geneva, Switzerland, 1999.

[18] N. Suresh, and N. Jayant, "'Mean Time Between Failures': A Subjectively Meaningful Video Quality Metric," in IEEE International Conference on Acoustics, Speech and Signal Processing, Toulouse, France, 2006.

[19] 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 Transactions on Consumer Electronics, vol. 56, no. 3, pp. 1311-19, 2010.

[20] G. Valenzise, S. Magni, M. Tagliasacchi, and S. Tubaro, "No-Reference Pixel Video Quality Monitoring of Channel-Induced Distortion," IEEE Transactions on Circuits and Systems for Video Technology, vol. 22, no. 4, pp. 605-618, 2012.

Author Biography

Jari Korhonen received his MSc in information engineering from University of Oulu, Finland (2001) and his PhD in telecommunications from Tampere University of Technology, Finland (2006). Since 2010, he has worked for the Department of Photonics Engineering at Technical University of Denmark. His current research interests focus on multimedia communications and visual quality assessment.

Claire Mantel received the M.S. and Ph.D. degrees in signal processing from Grenoble Polytechnic Institute, France, in 2007 and 2011, respectively. She is currently working as a researcher at the Department of Photonics Engineering of the Technical University of Denmark, Kongens Lyngby, Denmark. Her research interests include image and video coding and visual quality assessment.