Abstract

As video distribution over optical transport networks becomes more and more important, considerable focus is put on providing the appropriate quality for this transport service. A typical trade-off seen in this context is between sophisticated compression techniques used in order to save required transmission bandwidth, and the sensitivity of service quality with respect to transport network induced information (packet) loss. At the application layer, sensitivity to losses due to transport level impairments is content dependent, thus introducing a gap between typically used quality of service (QoS) parameters and the actual perceived quality (quality of experience – QoE). Extending a proposal to enhance transport level quality insurance mechanisms to be aware of the QoE significance of transmitted data, this paper presents a proposal to use significance information for QoE monitoring within the transport network as well. Two measurement techniques are described for delivering QoE indicators. Based on analytical expressions for the influence of packet loss on the MPEG-2 video decoding a Quality of Experience metric is derived which is applicable in significance and non-significance aware video transmission environments. The influence of different video encoding parameters on the overall achievable video quality and the expected gain in a significance aware environment are evaluated.

1 Introduction

Delivering considerable amount of multimedia content over the internet, like IP based Television broadcast (IPTV) and Video on Demand (VoD) services, has attracted increasing attention.

A particularly important aspect of audio/video transport is the quality of the delivered service. In contrast to classical data transport services, where some delay is accepted as a trade-off for lossless delivery, real-time streaming of audio/video data is much more sensitive to delay or delay variation, and requires quality insurance mechanisms compatible with such requirements. With native or overlay packet based networks used for delivering these services, unavoidable congestion situations have to be accounted for. Because audio/video transmission is sensitive even to low loss situations ([20]), it is worthwhile to assess the quality impact also of sporadic packet loss.

For video data, bandwidth demand for uncompressed transmission would be enormous. Increasingly sophisticated compression techniques are showing up, which all use the inherent redundancy of video data in the time dimension for data reduction. The established MPEG-2 standard ([5]), for instance, encodes only a relatively low number of video frames (pictures) internally (“I-Frames”). Most frames are only encoded as the delta information to preceding or also succeeding frames (“P-Frames”, “B-Frames”). As a consequence, any frame which serves as a reference for many delta-encoded frames (an I-Frame, in particular), is much more important on the receiving side than e.g. a B-Frame on which the decoding of no other frame is dependent. This fact introduces a differentiation within the transmitted data streams into packets which carry more significant (e.g. I-Frame) or less significant (e.g. B-Frame) information.

Complementing the idea to use this significance information within a ‘significance aware transport network’ to discard less significant packets first in the event of an error – thus providing graceful service degradation ([2],[3],[13]) – this paper presents the concept of using significance information also for monitoring purposes. This includes the basic mechanism for extracting the raw data, and an outlook how to derive ‘quality of experience’ measures analytically, based on those raw data.

2 Video stream structure overview

First, a short overview on the structure of compressed video data is given, and consequences for the significance of various parts of an audio/video data stream are pointed out.
2.1 Stream structure

A video stream is composed of successive picture frames. For MPEG-2 encoded video streams ([1],[5]), three basic picture frame types can be distinguished which are called I-, P- and B-Frames. The intra coded picture frames, called I-Frame, contain the information of transform coded pixel blocks of a single picture and provide only moderate compression. In P-Frames the picture information is predictively coded with respect to a reference frame. That means the coded information is based on information of the previous I- or P-Frame. Finally the B-Frames are bidirectional predictive coded. For decoding the B-Frame, information of the previous and following I- or P-Frame is required. B-Frames provide the highest information compression. One I-Frame and the subsequent P- and B-Frames till the next I-Frame are grouped together into a group of picture (GoP).

Figure 1: Sample MPEG-2 GOP Structure

Figure 1 shows the resulting picture decoding dependencies as well as the frame transmission order. Composing the video stream of independent GoPs reduces the error propagation in case of a picture frame loss.

Furthermore a picture frame can be subdivided into independently coded regions called slices. A slice itself is typically composed of a series of subsequent macro pixel blocks. This increases the error resilience, because if packets containing information of only one slice are lost, the rest of the picture can still be decoded. Therefore it might be sensible to map not more than one slice into a transport packet in order to prevent error propagation.

Although the considerations in this paper are mainly derived from MPEG-2 examples they are also applicable to more recent (and complex) techniques like H.264 or MPEG-4 video encodings ([16]). However this is beyond the scope of the paper and will not be presented here.

2.2 Significance

When it comes to rating the importance of the parts of an audio/video data stream, it has been demonstrated ([7]) that audio information is the most important of all; audio degradation is typically unacceptable, while some degree of degradation of the visual information for short times is often tolerated. So the audio part of a transported stream clearly has highest significance; on the other hand, this part does not need much bandwidth compared to the visual information and thus is less of a problem.

Focusing on video information, Figure 2 demonstrates the difference in impact of losing part of an I-Frame, P-Frame, or B-Frame (assuming the example GOP structure shown in Figure 1). Clearly, there is a drastic difference between losing an I-Frame or P-Frame, and losing a B-Frame.

Figure 2: Information Loss Impact per Frame Type

This implies that I-Frame information is most significant for the decoding process, P-Frame information slightly less, and B-Frame information is least significant.

Although the frame type, and thus the significance, is part of the encoded video data, this information is not readily available to equipment in the transport network. Inspecting the data stream on an application layer, and – at least partially – decoding the video stream in order to obtain that information, is just not economically feasible on that level. And even if it was, encryption techniques often used to prevent unauthorized access e.g. to a video broadcast service will make it technically impossible. Both issues can be addressed by passing the significance information to the network separately, within the overhead of the encapsulation used for transporting the stream. As an example, the real-time transport protocol (RTP) ([8]) is often used as an encapsulation for audio/video transport over IP based networks; significance marking could in this case be done for the RTP packets by adding this information to the RTP overhead.
3 Significance aware networks: ensuring and measuring QoE

A ‘significance aware’ network can now use the relative importance of transported data, as conveyed by an appropriate significance marking, to do ‘intelligent’ decisions on what parts of the data stream to drop and what parts to pass through in case of overload situations ([2],[3]). This would be an enhancement of the congestion control mechanisms already present in packet based or packet aware networks ([4]).

In addition, quality of experience degradation can be monitored within the network by introducing significance awareness monitoring points.

Making the network significance aware relies on a significance marking which is available to the respective network layer. The usual approach would be to pass down the significance level attribute from higher to lower layers through the adaptation between the layers, and pass it on within a layer by use of layer specific overhead information.

For example, significance could be encoded into IP header differentiated services (DS) fields ([10]), or Ethernet VLAN tags (P-Bits/DEI) ([6]). Alternatively, information could also be ‘snooped’ from a higher layer (similar to approaches like in [9]); this would on the one hand violate the strict layering, while on the other hand relax the need to make every layer be aware of significance information, can provide information at the application and the transport network significance aware (Figure 3). In any case, appropriate standardization would be required.

![Significance per Layer and Significance Snooping](image)

**Figure 3: Passing / Accessing Significance Information (Example)**

3.1 Significance and congestion handling

The primary goal of introducing significance information is to allow some control over the QoE impact in case of situations where information loss cannot be avoided (congestion). This can be achieved by using established traffic management strategies, use of “drop precedence” (or “drop eligibility”) aware congestion control mechanisms in particular ([12], [6]).

Significance levels can be mapped onto drop precedence levels (with higher drop precedence for lower significance), thus causing lower significance packets to be dropped first ([2], [3]).

3.2 Significance Indicators

For complementing significance aware congestion control, monitors can be placed into the network to provide indicators of QoE degradation between relevant demarcation points, like customer/provider or provider/provider interfaces.

Particularly relevant network performance parameters impacting audio/video QoE are packet loss and delay variation.

For both packet loss and delay variation, availability of significance information allows a more fine grained monitoring of these performance indicators, namely per significance level. In this way, a more accurate estimate of application level quality degradation can be derived.

4 QoE indicators: measurement

Measuring QoE indicators thus focuses on two types of audio/video transmission impairments: information loss during transport, and information loss at the receiving side due to exceeding acceptable delay or jitter.

Data stream packets may be lost in the network due to uncorrected bit errors, or due to packet switching induced losses, i.e., packet drop caused by congestion. The latter type is dominating, and is expected to be influenced by significance awareness. This means that one indicator to be monitored would be actual packet loss per significance level.

Data stream packets may also be lost at the receiver when late or early packets cannot be accommodated by the buffering capabilities of the decoding device, and are discarded due to overflow / underflow conditions. Monitoring packet delay, incorporating significance information, can provide information at the monitoring point about expected information loss in the receiving equipment due to discarded packets.

4.1 Information loss in the transport network

When packets are lost in the transport network, the information contained in the packet is not available any more for the decoding process. If a monitor just counts lost packets, it can not provide information on the induced QoE impact, because the perceived quality degradation will depend on the significance of the lost information.

Significance marking of packets provides information on QoE relevance of all received packets – if they are received, i.e., not lost. Having significance marked packets is not sufficient per se to allow counting lost
packets per significance level, as the marking is lost together with the packet. Some reference information is required to determine what exactly has been lost. Collection and correlation of monitor data in a network management system would not work: even leaving aside possible performance issues due to the amount of data to be transferred and processed, the involved time scales would not allow to achieve a precision for the data correlation which is appropriate for video frame frequencies of e.g. 25/second (and even higher frequencies for the transport level packets).

Our proposal is therefore to add reference data to the transmitted audio/video data stream by embedding additional loss measurement OAM (LM-OAM) control packets, analogous e.g. to the loss measurement concept described in ITU-T Y.1731 ([15]). These packets would be generated by upstream equipment – e.g. by the adaptation which is generating the significance marking of the audio/video data packets – and contain a count of transmitted packets, per significance level, up to the point where the LM-OAM packet is embedded into the data stream ([18]). With the information from the LM-OAM packets, plus local counters of received packets per significance level, packet loss counts per significance level can be derived (Figure 4). Sending the control data interleaved into the application data stream ensures exact timely correlation of local counters and received reference data.

![Figure 4: Significance Aware Packet Loss Monitoring](image)

### 4.2 Information loss caused by delay and jitter

Another type of information loss is not directly visible at the network equipment level as it occurs only in the decoder devices at the receiving end, after the audio/video data stream has left the network. Buffering capabilities in the decoder determine the degree of delay and jitter it can cope with before starting to discard packets.

Within the transport network, at least an estimate of these packet discards can be derived by introducing virtual buffers as monitoring devices ([18]). These will not actually buffer data stream packets, or even try to decode the audio/video data, but rather just emulate the respective buffering behavior. The monitor will maintain a running virtual buffer fill level, based on payload data sizes and application time stamp information. Figure 5 shows this buffer, which tracks relevant information per data packet. The virtual buffer will accept (virtual) packets up to a specified total length; entries are normally removed when a 'virtual decoder' consumes all virtual packets with a specific time stamp. (This process of course requires deeper packet inspection than the loss counting described previously, but can still be based on information available e.g. from an RTP encapsulation, without interpretation of the actual application data.) Discard events are counted in case of underflow – i.e. when packets arrive too late for the emulated decoding process to be available at the required presentation time (indicated by a dotted arrow to the left in Figure 5) -, or overflow – i.e. when packets arrive so early that there is no room in the buffer to keep them until they are needed for decoding (indicated by a dotted arrow to the right in Figure 5).

The virtual buffer monitor will track these discard events, together with the information (available from the significance marking) of the QoE impact of a discard: it collects (virtual) discard event counts per significance level.

![Figure 5: Virtual Buffer Based Discard Monitor](image)

### 4.3 Network deployment aspects

Overall goal of the significance aware monitoring is the measurement of QoE degradation. Depending on the exact responsibilities to be addressed, deployment of monitors could be done in several ways (Figure 6).

- At demarcation points between a content provider and the transport service provider, and between the transport service provider and the customer premises equipment (A)
- This allows determining if QoE degradation occurs within the transport network, or before entering / after leaving the network.
- At network-network interfaces (NNI) between different provider networks (B) Again, QoE degradation can be attributed to the part of the network where it occurs.
- Distributed within a single provider’s network (C) This would be done to allow for more fine grained localization of problems within a single responsibility domain.
- End-to-end, for monitoring overall service delivery This would be of minor relevance, because end-to-end monitoring is already available with established means like receive side monitoring and quality parameter feedback built into the RTP control protocol (RTCP) ([14])

Figure 6: Monitor Deployment Options

5 Deriving QoE measures from significance based indicators

In the following sections we assume that we only have rare packet loss events. That means that in a Group of Picture (GOP) there is at maximum one packet loss. Assuming just one packet loss per GOP simplifies the error modeling, as we do not have to take into account that some errors might conceal subsequent errors. E.g. if an I-Frame is lost it is not important whether the rest of the GOP is received correctly as through the differential encoding of the B- and P-Frames of the video data in a GOP the information can not be decoded anymore. The probability of a packet loss in a GOP depends on the overall packet loss probability as well as the number of packets in a GOP. Therefore in cases with a larger GOP structure, which means that there are more P- and B-Frames between two consecutive I-Frames, the probability of losing more than one packet in a GOP increases.

5.1 Rare loss event approximation

We assume for simplification a uniform packet loss distribution. This assumption might at first look relatively arbitrary as packet loss in the transport network is mainly induced due to buffer overflow and thus often shows a bursty character. While this is true for just one stream this significantly changes if we assume the parallel transport of several hundreds of video streams. A burst packet loss is now blurred over several video transport streams; thus the probability of losing more than one consecutive packet of one stream greatly decreases as the number of streams is increased. A uniform packet loss distribution seems also more valid if we assume a random early detection (RED) queuing behavior where packets are already dropped before the queue is full ([11]). The probability for having more than one packet loss per GOP can therefore be written as follows

\[ P(X \geq 2) = 1 - (1 - P_E)^N \cdot (N - 1) \cdot P_E + 1 \]

With \( P_E \) representing the packet loss probability and \( N \) the number of packets per GOP

Figure 7: Probability for more than one packet loss per GOP versus packet loss ratio

A standard video MPEG-2 data stream has a data rate between 3-4 Mbps and furthermore on average a ratio of 6:2:1 for the required amount of bandwidth between I, P and B frames can be assumed ([19]). That means an I-Frame requires three times the bandwidth of a P-Frame and 6 times more bandwidth than a B-Frame. We can see that for a video stream with 24fps, an average data rate of ~3.5Mbps and a reasonable GOP structure like IBBPBBPBB ~112 Ethernet packets per GoP would be required. This results with the above presented assumption in ~42 packets per I-Frame, ~14 packets per P-Frame and ~7 packets per B-Frame. For such a GOP structure and a packet loss rate of 10^-4 we have a probability of less than 0.01% that more than one packet error occurs in one GoP. This is shown in Figure 7.

5.2 Video Stream Quality Estimation

A very basic measurement for the estimation of the quality of a data stream is a simple count of the lost packets. The higher the packet loss ratio gets the lower the quality of the data stream will become.

\[ Q = 1 - P_E \]

Simply counting the packets however does not provide enough information to allow an estimate about the perceived video quality at the customer side. As
we do not have information about the transported video content, and also aim for an approach which is simple to implement, we concentrate our research on a reduced reference method ([17]). The only information we intend to use for the quality analysis are the significance markings and thus the frame type in the packet stream, including respective packet loss data. A first step would be to take the picture frame structure in the video transmission into account. Several packets might belong to one video frame thus the loss of one packet could lead to the loss of a whole picture frame. Furthermore we know that several picture frames are linked together in a group of picture.

Through the predictive differential encoding of the video data within a GoP a packet loss in one picture frame could influence several subsequent or preceding picture frames depending on the damaged frame type.

A packet loss in an I-Frame e.g. will have an impact on all other picture frames of the current GoP until the next I-Frame is received correctly. Thus the quality of the video stream can severely decrease with increasing packet loss rate depending on the affected frame type. In the following we assume that one packet loss leads to the loss of the whole picture frame and we estimate the video quality via the probabilities that the individual pictures frames are decodable ([3],[20]). The probability (and thus the quality) that an I-Frame can be decoded is consequently

\[ Q_I = (1 - P_E)^{M_I}, \]

where \( P_E \) denotes the packet loss rate and \( M_I \) represents the number of packets per I-Frame. The probability that a P frame in the GoP is decodable is the probability that the I-Frame and all previous P-Frames can be decoded and that the P-Frame itself is decodable as well. The aggregate quality for the collection of all P-Frames in a GoP can therefore be written as

\[ Q_P = \left(1 - P_E\right)\sum_{i=1}^{N_P} \left(1 - P_E\right)^{M_{P, i}}. \]

with \( M_P \) representing the number of packets of per P-Frame and \( N_P \) the number of P-Frames per GoP.

In a last step we define the quality for the B-Frames. Therefore we divide the GoP into several B-Frame blocks. A B-Frame block is always encapsulated by either two P-Frames or a P-Frame and an I-Frame. The number of B-Frames within a block depends on the GoP structure. The probability that a B-frame is decodable is dependent on the correct reconstruction of the immediately preceding and following I- or P-Frames. The aggregate quality for all B-Frames can be written as

\[ Q_B = N_B (1 - P_E)^{M_B} \sum_{i=1}^{N_{B,B}} (1 - P_E)^{M_B, i}. \]

with \( M_B \) reflecting the number of packets per B-Frame and \( N_B \) represents the number of B-Frames between two P Frames or a P- and an I-Frame.

The total achievable quality depending on the packet loss probability is consequently:

\[ Q_G = \frac{Q_I + Q_P + Q_B}{N_{GoP}}, \]

\( N_{GoP} \) represents the number of picture frames in a group of pictures.

If we assume now a significance marking of the I-, P- and B-Frames packets we have different packet loss ratios \( P_E \) for the different packet types. This has to be taken into account in the achievable quality calculation. In a transport network element, the packet loss ratio for the different significance levels can be gained from loss and throughput counters. Thus, an estimation of the overall quality of the video stream can be derived from those performance counters.

In normal (low loss) cases it can be assumed that if the network is significance aware it will only drop packets of the lowest significance level. With the above presented number of packets per I-, P- and B-Frame and a GoP structure of IBBPBBPBB the B-Frame packets require roughly \((6*7)/112 = 37.5\%\) percent of the overall data rate of the video stream. Thus, if others than B-Frame packets are dropped, then the network will be in a severe error state, and counteractions should be initiated. Thus it is sensible to assume that during normal working conditions only B-Frame packets are lost. Therefore the quality for I-, P-, B-Frames can be rewritten as follows:

\[ Q_I = 1; Q_P = N_P; \]

\[ Q_B = N_B \left(1 - P_{E,B}\right)^{M_B} \left(1 + N_P\right) \]

with \( P_{E,B} \) representing the packet loss rate for the B-Frames. \( P_{E,B} \) will be larger than in the non significance aware scenario in order to keep a constant overall packet loss ratio for a fair comparison.

\[ \text{Figure 8: Video Quality Estimation} \]

From Figure 8 it can be seen that the estimated quality based on the packet loss count only (as defined in the beginning of this subsection) overestimates the actual video quality compared to taking frame loss impact into account. We can also observe that the ex-
pected quality, which we can estimate through simple packet counting per significance level, is greatly increased in a significance aware packet drop environment. The difference between both approaches is dependent on the number of impacted video frames and thus the mapping into packets and the GoP length. In the above presented example we assumed an IBPBBPBB GoP structure with 42 Ethernet packets per I-Frame, 14 packets per P-Frame and 7 packets per B-Frame.

As stated in section 2.1, a picture frame is actually composed of different independent slices. A packet loss thus leads only to the loss of a percentage of the image and not the whole picture frame. However our quality analysis can be easily expanded to this approach. We simply regard the different slices as independent video streams and determine the quality of each slice stream individually. The total video quality can then be calculated from the slice stream quality estimation. The main difficulty in estimating the quality of the video stream lies within the estimation of the GoP and slice structure. As we can just monitor the packet loss we have to estimate the actual number of B-packets per B-Frame. The techniques for estimating the number of B-Frames as well as the analysis of the slice structure shall not be discussed in this paper. However it shall be noted that this is required for accurate quality estimation.

5.3 Video quality improvements through significance awareness

The quality of a video stream is often expressed via the PSNR value ([21]). The PSNR (Peak Signal to Noise Ratio) is defined as the rea ratio of the peak signal value (e.g. a 8 Bit luminance value) to the MSE (Mean Squared Error) between the original (reference) and the reconstructed (distorted) pixel value.

\[
PSNR = 10 \times \log_{10} \left( \frac{255^2}{MSE} \right)
\]

\[
MSE = \frac{\sum_{i=1}^{P_h} \sum_{j=1}^{P_v} (Ref(i,j) - Dist(i,j))^2}{P_h \times P_v}
\]

\[
\Delta PSNR = PSNR_{SA} - PSNR_{NSA} = 10 \times \log_{10} \left( \frac{MSE_{SA}}{MSE_{NSA}} \right)
\]

In the following we assume that the decoder does not use error concealment techniques thus that defect pixel are reconstructed as black spots and Dist(i,j) can be assumed as zero. As the actual reference pixel value is unknown to us we assume an average pixel value of 128 (8 Bit value) and a uniform distribution of luminance values in the pictures. We can see that the difference in the PSNR only depends on the difference in the number of lost macro-blocks N(Lost_MB) as all other parameters e.g. like the macro-block size represented via MBv and MBh are canceled out. Thus the only important parameter is the number of slices or the lost percentage of the picture frame.

\[
\Delta PSNR = 10 \times \log_{10} \left( \frac{N(Lost\_MB)_{SA}}{N(Lost\_MB)_{NSA}} \right) = 10 \times \log_{10} \left( \frac{1 - Q_{NSA}}{1 - Q_{SA}} \right)
\]

If we assume the above presented example where one frame is composed of one slice we get:

\[
\Delta PSNR = 10 \times \log_{10} \left( \frac{11 + 42}{112} + \frac{14}{112} + \frac{14}{112} + \frac{14 + 6}{112} \right) = 7.87 dB
\]

That means in such a scenario we would gain ~7.87dB in PSNR if we use significance aware packet dropping.

The number of lost slices in a packet loss event strongly depends on the slice mapping into transport packets as well as the used GoP length. The gain in PSNR increases with larger GoP structures. If we use a finer slice structure the gain will be constant as long as we assume that not more than one slice is mapped into a transport packet. However the overall video quality will increase.

\[
\begin{array}{|c|c|}
\hline
PSNR [dB] & MOS \\
\hline
>37 & 5 (Excellent) \\
31-37 & 4 (Good) \\
25-31 & 3 (Fair) \\
20-25 & 2 (Poor) \\
<20 & 1 (Bad) \\
\hline
\end{array}
\]

From Table 1 we can see that an increase of 7.87 dB can greatly improve the perceived video quality (Mean Opinion Score, MOS) for the user. In our example this corresponds to roughly one MOS level. Thus through introducing a significance aware packet dropping a better video quality can be guaranteed by the transport provider in an error case. Taking into account the slice structure within the video frames remains a subject for further research.
6 Summary

The paper presents an overview of the concept of significance differentiation in support of QoE insurance within the transport network domain, including the use of significance information to control packet drop decisions in congestion situations.

Complementing the method to ensure better application level quality, significance aware monitoring mechanisms have been introduced which allow to track expected QoE degradation at dedicated points in the network, based on actual (transport level) packet loss as well as 'virtual' expected packet loss due to delay and jitter.

Finally, an outline was given how the basic monitor data can be used to derive quality measures for the transported video streams, based on analytical expressions for quality as a function of packet loss rates per significance level. Using these analytical expressions, the difference between non significance aware and significance aware transport networks was illustrated.

7 Acknowledgment

This work has been partially funded by the German Ministry for Research and Education (BMBF Grant MUNAS 01BP554).

8 References