TOWARDS CONTENT-RELATED FEATURES FOR PARAMETRIC VIDEO QUALITY PREDICTION OF IPTV SERVICES.

M.N. Garcia, A. Raake

Quality and Usability Lab
Deutsche Telekom Laboratories,
Berlin University of Technology
Berlin, Germany

P. List

T-Systems Entreprises Services GmbH
Darmstadt,Germany

ABSTRACT

This paper investigates video content-related features, such as measures of spatio-temporal complexity, for inclusion into parametric video quality models. Our goal is to find a parametric content description that correlates with perceived video quality. In the course of the development of a parametric IPTV video quality prediction model (T-V-Model [1]), a large number of subjective tests have been conducted for Standard Definition and High Definition video with different types of content. As expected from previous studies [2], we observed content dependencies that were different for different types of degradations. As descriptors of the content, we employ spatio-temporal related information obtained either before encoding and from the decoder or obtained from the decoder only. We compare those two approaches and explore their application to a reduced- or no-reference parametric model. An outlook highlights future steps for integrating the spatio-temporal features into the parametric model.

Index Terms— Video quality, video content, spatio-temporal complexity, parametric-bitstream model, IPTV

1. INTRODUCTION

The influence of the content on perceived video quality is under study in many applications like the selection of test material for video quality assessment [2], video segmentation,video classification [3] and quality prediction [4],[5],[6]. We started working on this content-issue when developing a parametric video quality prediction model for IPTV services. This model provides estimates of the video quality as perceived by the user on the basis of a parametric description of the video end-to-end transport and processing path, as shown in Figure 1. This model can be used for planning IPTV networks and IPTV service quality monitoring. The results of the subjective tests conducted for developing this model clearly reflect a variation of the perceived video quality in function of the content. This variation is described in Section 2 along with the video contents used in our tests. The influence of the content is strongly related to the difficulties the encoder encounters when encoding video sequences with a large amount of details, complex structures and complex movements, also called spatio-temporal information of the video sequence. We present in Section 3 the spatio-temporal features we extract from the video sequences as descriptors of the influence of the content on the perceived quality. As shown in figure 1, two approaches are followed for extracting the spatio-temporal features: the "reduced-reference" approach in which we have access both to a reduced version of the original signal (before encoding, at the sender side) and the bitstream (at the receiver side), and the "no-reference" approach in which we have access only to the bitstream, which is the case when measuring and monitoring the video quality of video IP services. We derive measures from those spatio-temporal features and, in Section 4, study the adequacy of the measures with the perceived video quality. In addition, we analyze the benefits of accessing the original signal ("reduced-reference" approach) over a network- or client-based approach.

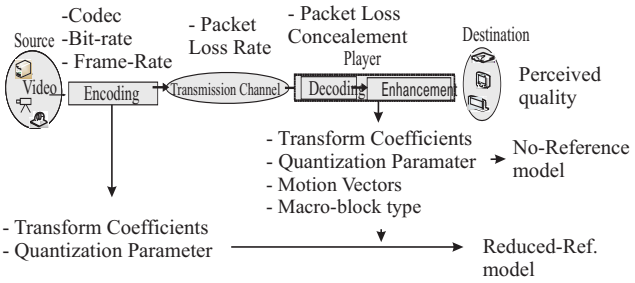

Fig. 1. Parametric model and extraction of content-based features.

2. IMPACT OF CONTENT ON VIDEO QUALITY

As shown in figure 1,the T-V-Model takes as input parameters such as the type of codec, the target bit-rate, the packet loss

rate and the packet loss concealment type. Those parameters

Parameters	Values
Codecs	MPEG2 and H264
Bit-rates	2 Mbps to 64 Mbps
Packet Loss Rates	0% to 4%
Packet Loss Concealment	Freezing and Slicing

Table 1. Test Parameters.

were applied on five video contents of 16s duration each and the so called "Absolute Rating" was used for the collection of subjective judgments on the quality of the presented video segments. This method is derived from the standardized "Absolute Category Rating" (ACR) method [7] [8]. The subjects judged the quality using an 11-point discrete quality scale [8].

The five video contents used for the tests and listed in table 2 are representative of various TV programs and with different amount of details and complexity of structures and movements. Figure 2 shows the subjective judgments of the

ID	Category and Description
A	Movie Trailer: High amount of details, scene cuts and movements, fast panning, presence of explosions, night light
B	Interview: Close-up shot, panning,indoor
C	Soccer: High amount of details, complex movements, complex structure, outdoor
D	Movie: High amount of details, complex movements, slow panning, zoom, day light
E	Music Video: Medium amount of details and complex movements Presence of scene cuts

Table 2. Content Description.

video sequences depending on the video content and the type of degradation. We observe that apart from the lowest bit-rate, the movie (magenta square) and the interview (green star) obtain the best rating in terms of quality while the movie trailer (red star) and the video clip (cyan triangle) yield equal ratings. The ranking of the soccer video improves when the bit-rate increases. This soccer video seems to be really sensitive to the bit-rate and requires high bit-rate encoding for getting an acceptable quality. The ranking of the contents for the lowest bit-rate is clearly different from the ranking of the contents from medium and high bit-rates.

Based on these observation, it became clear that we had to introduce parameters in our video quality model which would modulate the predicted video quality as a function of the content. This starts with finding appropriate descriptors of the

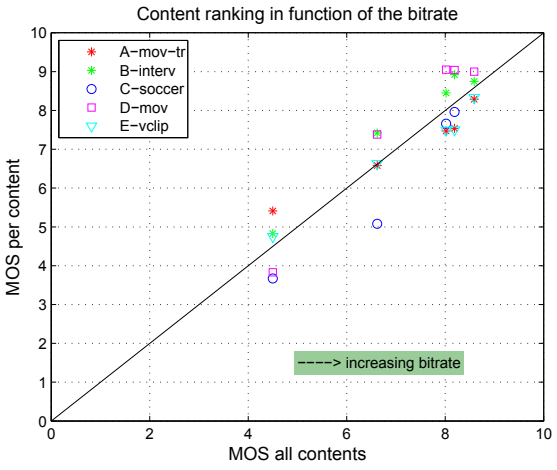

Fig. 2. Perceived quality depending on the bit-rate and on the content.

content. Appropriate in our case means both representative of what the encoder considers as complex to encode and, since our model has access only to the signal at the receiver side, descriptors that can be extracted from the bitstream. As information is lost in case of lower bit-rate coding, it may not be available at a network or client-site measurement point. Hence, we also analyze the extraction of spatio-temporal descriptors from the original. As a first step, we will concentrate our analysis on High Definition video sequences and H.264 encoding.

3. CONTENT DESCRIPTORS

3.1. Spatial features

When encoding the video sequence, the H.264 encoder [9] transforms each 4x4 block into 16 transform coefficients using an integer transform. The first coefficient is the DC frequency component of the signal and the other 15 transform coefficients are the AC frequency components of the signal at various horizontal and vertical frequencies. High frequency AC components in I-Frame indicate the presence of details and complex structure, i.e high spatial complexity. The quan-

Spatial complexity measures
The average of the 15 AC coeff. of each macro-block averaged over each I-frame and over the whole video.
The standard deviation over the 15 AC coefficients of each I-frame averaged over the whole video sequence.
The Quantization Parameter (QP) averaged over each I-frame and over the whole video sequence.

tization parameter (QP) is a measure for the overall quality. Indeed, every rate control mechanism adjusts the QP in such

a way that the given data-rate is kept constant regardless of how high or low the temporal/spacial complexity or picture size is. Thus, at least for I-frames, a certain QP means a loss of detail, which interacts with the amount of detail present in the original video.

3.2. Temporal features

Very high temporal complexity in a coding sense usually occurs when many small objects move chaotically. In that case the standard deviations of the horizontal and vertical components of the motion vectors are high.

Each P- and B- frames can contain different types of macro-blocks:

Skipped-Macro-blocks: Usually, a macro-block becomes "Skipped" when its prediction with the default motion vector (the predicted motion vector) is good. In this case nothing needs to be transmitted. Macro-blocks with motion-shape 16: The macro-block is represented with only one motion vector per (16x16) macro-block. This macro-block is easy to predict and therefore has a low temporal complexity. Macro-blocks with motion-shape 8 and 4: More vectors are needed per macro-block which means that blocks are more complex to predict. Intra-macro-block within a P-frame is always a sign that the scene could not be predicted well. This is for instance the case if the encoder has not employed scene cut detection. The frame after the cut will automatically contain a lot of intra-macro-blocks.

For all those reasons, we measure as descriptors of temporal complexity:

Temporal complexity measures
The amplitude of the motion vectors averaged over all P- and B-frames and over the whole video sequence.
The standard deviation of the horizontal and vertical components of the motion vectors, averaged over all P- and B-frames and over the whole video seq.
For each macro-block type, the number of macro-blocks averaged on all P- and B-frames and over the whole video sequence.

In H.264, motion vectors appear as 16*16, 8*16, 16*8, 8*8, 4*8, 8*4 and 4*4 vectors. We normalized them on 4*4 vectors (one 16*16 vector becomes 16 4*4 vectors).

For the reason of simplicity, motion vectors of B-Frames are treated as in the case of P-frames (only the forward aspect of their motion is taken into account).

3.3. Reduced-Reference content descriptors

We want to know how much information we loose by measuring the spatio-temporal features at the receiver side only instead of using information extracted at the sender side. There are two ways in which information can be lost: the packet losses occurring in the network and the loss linked to the H.264 encoding process. In this paper, we consider only the second case, which results in analyzing the loss of high frequency transform coefficients due to the compression (and especially to the quantization of those coefficients). To do so we need to know the transform coefficients in case there was no coding. We simulated this ideal case by extracting the transform coefficients of a highly encoded signal. Then, we do the same calculation as for the spatial features in section 3.1.

4. ADEQUACY OF QUALITY AND CONTENT DESCRIPTORS

4.1. Spatial features

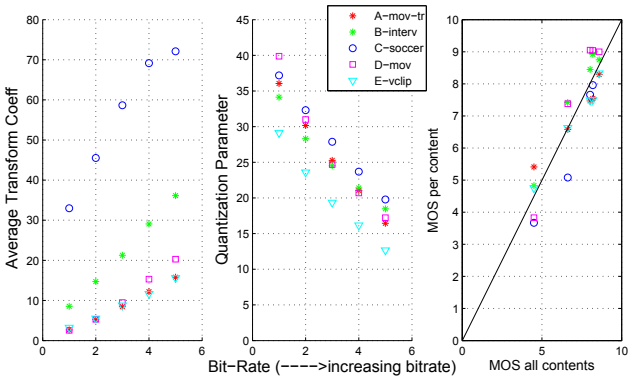

Fig. 3. Spatial Complexity.

Those figures clearly separate the soccer video from the other video sequences. Indeed,the average of the 15 AC transform coefficients are much higher for the soccer video than for all the other video sequences. Moreover, except for the lowest bit-rate, the quantization parameter of the soccer video is also higher than for the other videos. This means that the soccer video contains high frequency transform coefficients which are lost due to a coarse quantization. Since the bit-rate is mainly controlled by the quantization parameter (QP), this parameter increases when the bit-rate is reduced. There is a noticeable difference between the QP's of the different sequences. In particular the music clip seems to be the easiest to code.

The exceptionally high quantization parameter extracted from the movie bitstream at low bit-rate could be an explanation of its low rating, despite its low average transform coefficient.

It is noticeable that the movie trailer and the video clip have the same averaged transform coefficients, whatever the bit-rate. This similarity is also observed for some of the temporal features and is in agreement with the close ratings between the two video sequences.

At last, except for the lowest bit-rate for the quantization parameter, the ranking of the contents by bit-rate is conserved among the bit-rates. As a consequence, extracting our spatial features at really high bit-rate, which simulates the ideal case where we have access to the original signal, does not bring additional information compared to the case where we extract the features at the receiver side only.

4.2. Temporal features

As expected, the standard deviations of the X and Y components of the motion vectors do not vary in function of the bit-rate. Indeed, those two standard deviations represent how "chaotic" the movements of the video scenes are. This measure of the chaotic aspect is not affected by the bit-rate.

The movie presents the highest averages for the standard deviation of the X and Y components of the motion vectors while the averages for the soccer are the lowest.

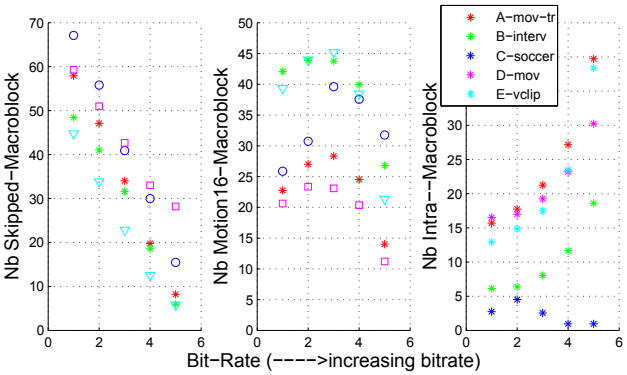

Fig. 4. Statistics on Macro-block types.

As in the case of the spatial complexity:

- The ranking of the content is independent of the bit-rate.
- The movie trailer and the music video either are neighbors in the ranking or have similar values for their features.
- The values of the extracted features for the soccer video are extreme values.

5. CONCLUSION AND OUTLOOKS

This paper presents a pilot research for predicting the perceived video quality in function of the content. At this stage, the proposed spatial and temporal features do not quantitatively predict the perceived quality in function of the content but already highlight a strong adequation with the quality. Indeed, it has been for instance observed that contents which have similar quality rating have features with similar values, which is extremely encouraging.

From this point, several steps have to be completed:

- Study the combination of spatial and temporal features for predicting the perceived video quality in function of the content,
- Extend our analysis on a large variety and number of contents,
- Extend our analysis on the prediction of the perceived video quality in function of the packet loss and the content,
- Integrate the spatio-temporal descriptors into the parametric model and study the benefit of this integration. This would lead to a hybrid parametric/bitstream model in the scope of P.NAMS, which is a standard currently developed by Study Group 12 of ITU-T (Non-intrusive parametric model for the Assessment of performance of Multimedia Streaming),
- Investigate other measurements for the reduced-reference spatial complexity (wavelet transforms, ITU measurements [2]),
- and conduct future studies for associating content-related features with information extracted from visual attention models and semantic information.

6. REFERENCES

[1] A. Raake, M.N. Garcia, S.Moeller, J. Berger, F. Kling, P. List, J. Johann, and C. Heidemann, "T-V-MODEL: Parameter-based prediction of IPTV quality," in *ICASSP08, in process*, 2008.

[2] "ITU-R recommendation BT.1210: Test materials to be used in subjective assessment," 2004.

[3] Y. Liu, R. Kurceren, and U. Budhia, "Video classification for video quality prediction," in *Journal of Zhejiang University Science A*, 2006.

[4] H. Koumaras, A. Kourtis, D. Markatos, and J. Lauterjing, "Quantified PQoS assessment based on fast estimation of the spatial and temporal activity level," in *Springer Science*, 2007.

[5] M. Ries, C. Crespi, O. Nemethova, and M. Rupp, "Content-based Video Quality Estimation for H.264/AVC Video Streaming," in *IEEE Wireless Communications and Networking Conference*, 2007.

[6] S. Wolf and M.H. Pinson, "Spatial-temporal distortion metric for in-service quality monitoring of any digital video system," in *Proceedings of SPIE,Multimedia Systems and Applications II*, 1999, vol. 3845, pp. 266–277.

[7] "ITU-R recommendation BT.500-11:Methodology for the subjective assessment of the quality of television pictures," .

[8] "ITU-T recommendation P.910: Subjective video quality assessment methods for multimedia applications," 1999.

[9] T. Wiegand, G.J. Sullivan, G. Bjontegaard, and A. Luthra., "Overview of the H.264/AVC Video Coding Standard," in *IEEE Trans. on Circuits and Systems for Video Technology*, 2003.