Motion Based Reference-Free Quality Estimation for H.264/AVC Video Streaming

Michal Ries, Olivia Nemethova and Markus Rupp

Institute of Communications and Radio-Frequency Engineering
Vienna University of Technology
Gusshausstasse. 25, A-1040 Vienna, Austria
(mries, onemeth, mrupp)@nt.tuwien.ac.at

***Abstract*— The scope of this paper is the estimation of subjective video quality for low resolution video sequences as they are typical for mobile video streaming. Although the video quality experienced by users depends on spatial (edges, colors, ...) and more considerably on temporal (movement speed, direction, ...) features of the video sequence, most of the proposed methods are based on spatial features. This paper presents a new reference free approach for quality estimation based on motion character. The character of motion is determined by the amount and direction of the motion between two scene changes. In this paper, the design of a universal reference-free video quality metric is presented. The design of the quality metric is based on content adaptive parameters, allowing for content dependent video quality estimation. The performance of the proposed universal metric was evaluated with various video content types and compressions settings. The results show that the proposed approach provides powerful means of estimating the video quality experienced by users for low resolution video streaming services.**

I. Introduction

For the provisioning of video streaming services it is essential to provide a required level of customer satisfaction, given by the perceived video stream quality. It is therefore important to choose the compression parameters as well as the network settings so that they maximize the end-user quality. Due to video compression improvement of the newest video coding standard H.264/AVC allows for providing video streaming for low bit and frame rates while preserving the perceptual quality. This is especially suitable for video applications in 3G wireless networks.

Mobile video streaming is characterized by low resolutions, and low bit-rates. The commonly used resolutions are *Quarter Common Intermediate Format* (QCIF,176x144 pixels) for cell phones, *Common Intermediate Format* (CIF, 352x288 pixels) and *Standard Interchange Format* (SIF, 320x240 pixels) for data-cards and palmtops (PDA). The mandatory codec for UMTS (Universal Mobile Telecommunications System) streaming applications is H.263 but the 3GPP release 6 [1] already supports a baseline profile of the new H.264/AVC codec [2]. The appropriate encoder settings for UMTS streaming services differ for different streaming content types [3], [4], [5], [6] and streaming application settings (resolution, frame and bit rate).

In the last years, several objective metrics for perceptual video quality estimation were proposed. The proposed metrics can be subdivided into two main groups: human vision model based video metrics [7], [8], [9], [10] and metrics based only on the objective video parameters [11], [12], [13], [14]. The complexity of these methods is quite high and they are mostly based on spatial features, although temporal features reflect better subjective quality. Most of these metrics were designed for broadband broadcasting video services and do not consider mobile video streaming scenarios. Moreover, we are looking at measures that do not need the original (non-compressed) sequence for the estimation of quality, because this reduces the complexity and at the same time broadens the possibilities of the quality prediction deployment. Hence, we are looking for an objective measure of the video quality simple enough to be calculated in real-time at the receiver side.

The goal of our research is to estimate the video quality of mobile video streaming at the user-level (perceptual quality of service) for any possible codec settings in 3G network and for any content type.

The paper is organized as follows: In Section 2 we describe test setup for video quality evaluation. In Section 3 the process of motion features extraction is explained. The results are introduced and further processed in Section 4, where the focus is given on the video quality estimation. Section 5 contains conclusions and some future work

II. The test setup for video quality evaluation

For the tests we selected two sets of five video sequences each having ten second duration, SIF resolution and encoded with an H.264/AVC baseline profile 1b. We choose five most frequent contents with different impact on the user perception. In the ”news” sequences a moderator is reading news only

Fig. 1. Snapshots of typical news content

by moving her lips and eyes (see Figure 1). The ”news” sequences include sequences with a small moving region of interest (face) on static background. The ”soccer” sequences contain wide angle camera sequences with uniform camera movement (panning). The camera is tracking a small rapid

Fig. 2. Snapshots of typical soccer content

moving object (ball) on the uniformly colored (typically green) background (see Figure 2).

Fig. 3. Snapshots of typical cartoon content

1) Content class number three (cartoon): In "cartoon" sequences object motion is dominant, background is usually static (see Figure 3). The global motion is almost not present due to the artificial origin of the movies (no camera). The object movement has no natural character. "Panorama" se-

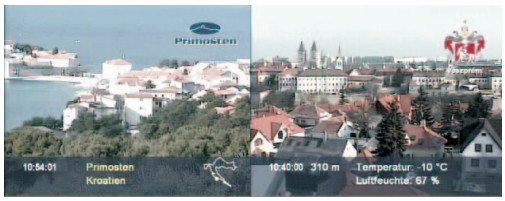

Fig. 4. Snapshots of typical panorama content

quences contain global motion sequences taken with wide angle panning camera (see Figure 4). The camera movement is uniform and in one direction. The last investigated sequence

Fig. 5. Snapshots of typical video clip content

videoclip contains a lot of global and local motion or fast scene changes (see Figure 5).

A. Test setup

For subjective quality testing we used typical video codec settings (see Table I) for mobile video streaming services. In total they were 36 combinations, but we excluded some combinations where the resulting video quality was clearly insufficient.

FR [fps]/BR [kbit/s]	24	50	56
5	Ne, Ca	Vi	Ne, Ca
7.5	Ne, Ca		Ne, Ca
10	Ne, Ca		Ne, Ca
15	Ne		Ne

FR [fps]/BR [kbit/s]	60	70	80	105
5				Ne
7.5	Vi	Vi		Ne, So, Vi
10		Vi	Vi	Ne, So, Vi
15			Vi	Ne, So, Vi

TABLE I

TESTED COMBINATIONS OF FRAME RATES AND BIT RATES. ABREVIATION OF SEQUENCE TYPES: CA = CARTOON, NE = NEWS, SO = SOCCER, PA = PANORAMA, VI = VIDEOCLIP

To obtain a MOS (Mean Opinion Score), we worked with 36 test persons (the training set with 26 and the evaluation set with 10 persons) for two different sets of test sequence. The the training and evaluation tests were collected of different sets of five video sequences. The tests were consistent with the ITU-T Recommendation [15], using absolute category rating (ACR) method as it better imitates the real world streaming scenario. We did not follow ITU-T Recommendation [15] only in one case and in order to emulate real conditions of the UMTS service [4], all the sequences were displayed on a PDA VPA IV UMTS/WLAN (see Figure 6).

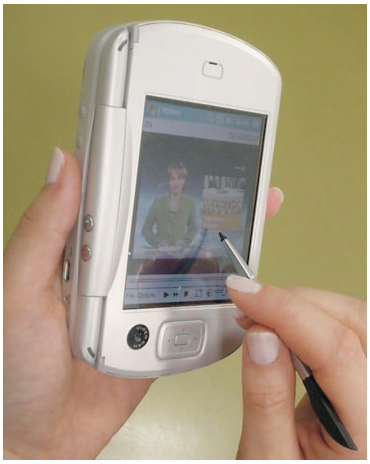

Fig. 6. Test equipment: VPA IV UMTS/WLAN

III. FEATURES EXTRACTION

The human visual perception of video content is determined by the character of the observed sequence. The character of a sequence can be described by spatial information [13], [14]. Such approaches come mainly from the quality estimation of still images [16], [17]. Equivalently, the motion characteristics can be used to characterize the sequence. In small resolutions and after applying compression, not only speed of movement (influencing at most the compression rate) but also the type

and direction of movement (temporal information) play an important role in the user perception. Therefore, in this work we focus on the motion features of the video sequences that determine the perceived quality.

1) Temporal segmentation: Since the sequence can contain different scenes-shots with different characteristics, we segment each sequence first by a scene change detection based on a dynamic threshold [18]. For our purpose the method was improved, extended to all content types.
The thresholding function is based on a local sequence statistical features. The higher accuracy was reached by introducing 10 forecoming and 10 upcoming frames into averaging. We calculate a sum of absolute differences (SAD) between two frames (n and $n+1$). Moreover the following local statistics are computed, empirical mean m_n and standard deviation σ_n for a sliding window $[n-10, n+10]$:

$$\mathrm{m}_n = \frac{1}{2N} \sum_{n-N}^{n+N} \mathrm{SAD}_n \tag{1}$$

and

$$\sigma_n = \sqrt{\frac{1}{2N-1} \sum_{n-N}^{n+N} (\mathrm{SAD}_n - m_n)^2}. \tag{2}$$

The equations (1) and (2) are used for defining the variable threshold function:

$$\mathrm{T}_n = a \cdot m_n + b \cdot \sigma_n. \tag{3}$$

The constants a, b were tuned up in order to get the best performance for all content types. The constant a was set in order to avoid wrong change detections like in case of intense motion scenes; but on the other hand, the detector can miss some low valued, difficult scene changes. The b constant was tuned in order to prevent from detecting the intense motion as a scene change as you can see in Figure 7. The scene change detector works with precision and recall higher than 97%.

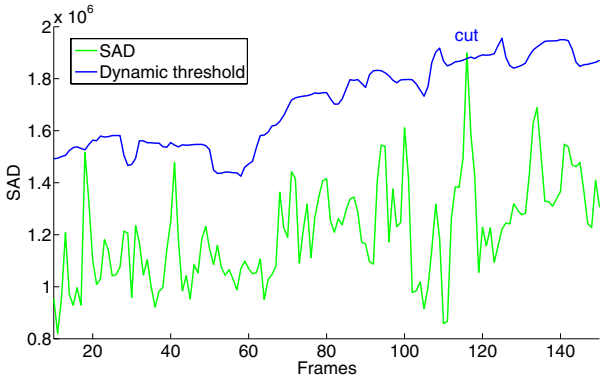

Fig. 7. Performance of dynamics threshold function on a sequence with global rapid movement (car race).

2) Extraction of sequence motion parameters: The static or dynamic character of a sequence is one of the main reasons for the differences in perceived quality. That is why we investigated MV features. The MVs were calculated for block size of 8×8 pixels. What divides a QVGA screen (320×240 pixels) into the 30×40 blocks (1200 MVs per frame). Further investigation of the MV bloc size shows that increasing block size leads to the significant lost of motion information. On the contrary decreasing block size results to the rapid increase of processing power and makes the MV estimation sensible on invisible changes which do not affect the human vision.
The size distribution and the directional features of the MVs were analyzed within one the sequence between two cuts. Furthermore zero MV vector allows us for estimating size of still region. That allows to analyze MV features separately for region with movement. This particular MV features make possible to detect rapid local movements or character of global movement. We investigated following statistical MV features with and without still region:

- mean size of all MV
- standard deviation of MV sizes
- histograms of MV directions
- variance of MV directions
- proportion of horizontal movement
- proportion of dominant MV direction

In total 12 MV features, bit rate (BR) and frame rate (FR) were calculated. Furthermore, it was necessary to investigate influence of of these motion parameters and the BR and the FR on investigated content.
For this purpose, we used a well known multivariate statistical method, the Principal Component Analysis (PCA) [20]. The PCA was carried out to verify further applicability of the motion characteristics, BR for metric design. In our case first two components proved to be sufficient for an adequate modeling of the variance of the data. The variability of the first component is 42.1% and second 20.6%. The PCA results (see Figure 8) show sufficient influence of most significant parameters on our data set for all content classes.
Following MV features and BR to represent the motion characteristics:

- **Zero MV ratio within one shot Z:**
  Percentage of zero MVs within one shot. It is a proportion of the frame that does not change at all (or changes very slightly) between two consecutive frames averaged over all frames in the shot. This feature detects the proportion of still region. The high proportion of the still region refers to very static sequence with small significant local movement. The viewer attention is focused mainly on this small moving region. The low proportion of still region indicates uniform global movement and/or a lot of local movement.
- **Mean MV size within one shot N:**
  Proportion of mean size of the non-zero MVs within one shot normalized to the screen width, expressed in percentage. This parameter determines intensity of movement within moving region. Low intensity within large moving region indicates that importance of static quality. High intensity within large moving region indicates rapidly changing scene.

- **Ratio of MV deviation within one shot S:**
  Proportion of standard MV deviation within one shot to mean MV size within one shot, expressed in percentage. High deviation indicates a lot of local movement and low deviation indicates global movement.
- **Uniformity of movement within one shot U:**
  Percentage of MVs pointing in the dominant direction (the most frequent direction of MVs) within one shot. For this purpose, the resolution of the direction is 10^o. This feature express proportion of uniform and local movement within one sequence.
- **Average BR:**
  Refers to pure video payload. The BR is calculated as an average over the whole stream. BR reflect compression gain in spatial and temporal domain. Moreover the encoder performance is dependent on the motion characteristics. The BR reduction causes loss of the spatial and temporal information what is usually annoying for viewers.

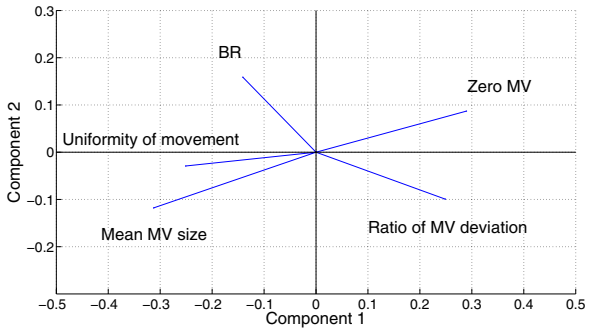

Fig. 8. Visualization of PCA results for all content classes.

The perceptual quality reduction in spatial and temporal domain is very sensitive to the chosen motion features. Thats makes motion features very suitable for reference free quality estimation because higher compression does not necessarily reduce subjective video quality (e. g. in static sequences).

IV. Video quality estimation

The subjective video quality is estimated with five objective parameters. Additional investigated objective parameters do not improve the estimation performance. On the other hand reducing of objective parameters decrease significantly estimation accuracy. The proposed model reflects relation of objective parameters to MOS. Furthermore the mix-term show mutual dependence of the movement intensity and its character (global or local movement). Finally we propose one universal metric for all contents based on the defined motion parameters Z, S, N, U and the BR:

$$\begin{aligned}\widehat{\mathrm{MOS}} = \quad & a + b \cdot \mathrm{BR} + c \cdot \mathrm{Z} + d \cdot \mathrm{S}^e + \\ & + f \cdot \mathrm{N}^2 + g \cdot \ln(\mathrm{U}) + h \cdot \mathrm{S} \cdot \mathrm{N}.\end{aligned} \tag{4}$$

The metric coefficients were obtained with a regression of the proposed model with our training set (MOS values averaged over two runs of all 26 subjective evaluations for particular test sequence). To evaluate the quality of the fit of our proposed metrics for our data, we used a Pearson correlation factor [19]. The metric model was evaluated with a MOS values from evaluation set (MOS values averaged over two runs of all 10 subjective evaluations for particular test sequence).

Coeff.	
a	4.631
b	8.966×10^{-3}
c	8.900×10^{-3}
d	-5.914×10^{-2}
e	0.783
f	-0.455
g	-5.272×10^{-2}
h	8.441×10^{-3}

TABLE II
Coefficients of metric model

Metric	Pearson correlation [%]
Motion based	80.25
Content based	81.93
ANSI	41.73

TABLE III
Coefficients of metric model

Furthermore we compare proposed metric with content based metric [21] and well known ANSI metric [11]. The content and motion based metrics are proposed for mobile streaming, the ANSI metric is proposed as a universal prediction metric. The motion based metric is full reference-free estimator. Content based metric is reference free estimator with additional content classification [21]. Finally, the ANSI is reference metric. The obtained prediction performance on evaluation set (see Table III and Figure 9) shows good agreement between between MOS and estimated MOS results. The weak performance of ANSI shows that this metric is not suitable for mobile streaming scenario. The usage of the mobile streaming services influence the subjective evaluation. Therefore the universal metrics are not suitable for estimation of mobile video quality. The detailed investigation shows that ANSI metric has better prediction performance for higher MOS values. This explains our observation that ANSI has better performance for bitrates over 90 kbps.

V. Conclusion

In this paper we propose a motion based video quality metric for mobile video streaming services. The comparison with the reference universal ANSI metric shows the relevance of our approach. The universal method is not suitable for video quality estimation due to its complexity and different perception of the mobile video streaming services. Furthermore, the proposed method allows us for continuous and reference free quality measurements on both transmitter and receiver side. This feature extends the applicability of the proposed method. Moreover, we reveal a mutual relation of the content character, motion features and subjective video quality. This allows us for good estimation on contents with significantly different content characters.

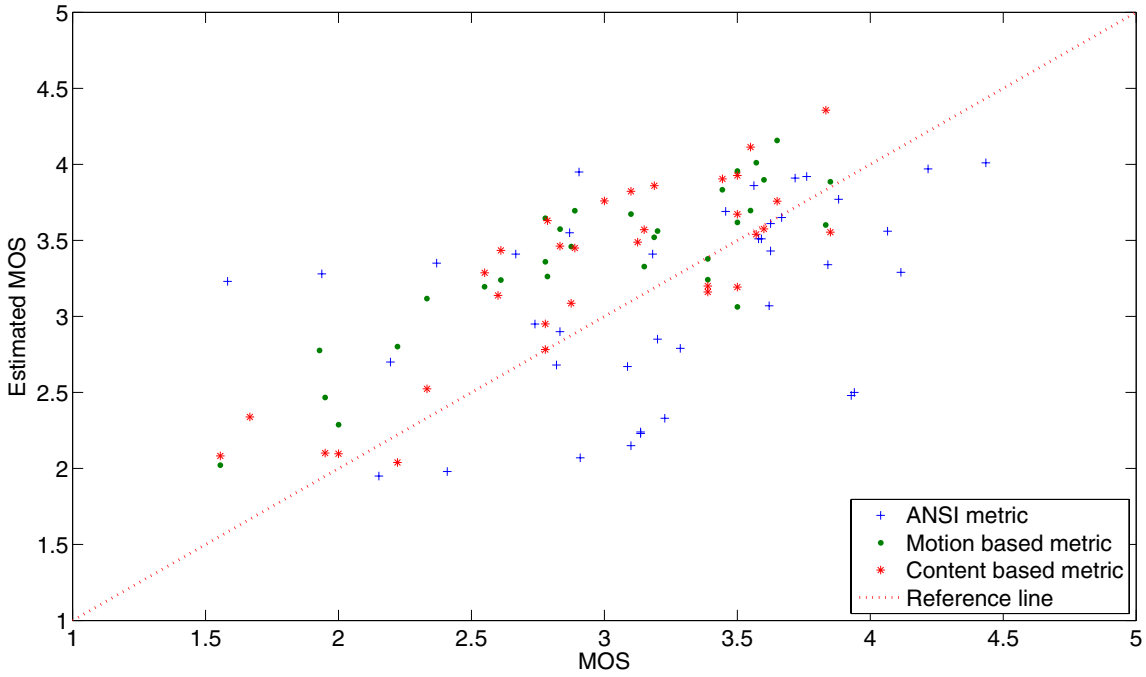

Fig. 9. Estimated vs. subjective MOS results

VI. Acknowledgment

The authors would like to thank mobilkom austria AG for supporting their research. The views expressed in this paper are those of the authors and do not necessarily reflect the views within mobilkom austria AG.

References

[1] 3GPP TS 26.234 V6.8.0:"End-to-end transparent streaming service; Protocols and codecs".

[2] ITU-T Recommendation H.264 (03/05): "Advanced video coding for generic audiovisual services" — ISO/IEC 14496-10:2005: "Information technology - Coding of audio-visual objects - Part 10: Advanced Video Coding".

[3] M. Ries, O. Nemethova, M. Rupp. "Reference-Free Video Quality Metric for Mobile Streaming Applications," Proc. of the DSPCS 05 & WITSP 05, pp. 98-103, Sunshine Coast, Australia, December, 2005.

[4] O. Nemethova, M. Ries, E. Siffel, M. Rupp, "Quality Assessment for H.264 Coded Low-Rate and low-Resolution Video Sequences," Proc. of Conf. on Internet and Inf. Technologies (CIIT), St. Thomas, US Virgin Islands, pp. 136-140, 2004.

[5] H. Koumaras, A. Kourtis, D. Martakos, "Evaluation of Video Quality Based on Objectively Estimated Metric," Journal of Communications and Networking, Korean Institute of Communications Sciences (KICS), vol. 7, No.3, Sep. 2005,

[6] C. John, "Effect of content on perceived video quality," Univ. of Colorado, Interdisciplinary Telecommunications Program: TLEN 5380 Video Technology, 9 Aug. 2006

[7] A. W. Rix, A. Bourret, and M. P. Hollier, "Models of Human Perception," J. of BT Tech., vol. 17, no. 1, pp. 24-34, Jan. 1999.

[8] S. Winkler, F. Dufaux, "Video Quality Evaluation for Mobile Applications," Proc. of SPIE Conference on Visual Communications and Image Processing, Lugano, Switzerland, vol. 5150, pp. 593-603, July 2003.

[9] S. Winkler, Digital Video Quality, JohnWiley & Sons, Chichester, 2005.

[10] E.P. Ong, W. Lin, Z. Lu, S. Yao, X. Yang, F. Moschetti, "Low bit rate quality assessment based on perceptual characteristics," Proc. of Int. Conf. on Image Processing , vol. 3, pp. 182-192, Sep. 2003.

[11] ANSI T1.801.03, "American National Standard for Telecommunications - Digital transport of one-way video signals. Parameters for objective performance assessment," American National Standars Institute, 2003.

[12] M.H. Pinson, S. Wolf, "A new standardized method for objectively measuring video quality," IEEE Transactions on broadcasting, vol. 50, issue: 3, pp. 312-322, Sep. 2004.

[13] T. M. Kusuma, H. J. Zepernick, M. Caldera; "On the Development of a Reduced-Reference Perceptual Image Quality Metric," Proc. of the ICW05, pp. 178-184, Montreal, Canada, August, 2005.

[14] P. Marziliano, F. Dufaux, S. Winkler, and T. Ebrahimi, "A No-Reference Perceptual Blur Metric," Proc. of the IEEE Int. Conf. on Image Processing, pp. 57-60, Sep. 2002.

[15] ITU-T Recommendation P.910, "Subjective video quality assessment methods for multimedia applications," Sep. 1999.

[16] Z. Wang, H. R. Sheikh, and A. C. Bovik, "No-Reference Perceptual Quality Assessment of JPEG Compressed Images," Proc. of the IEEE Int. Conf. on Image Processing, pp. 477-480, Sep. 2002.

[17] S. Saha and R. Vemuri, "An Analysis on theEffect of Image Features on Lossy Coding Performance," IEEE Signal Processing Letter, vol. 7, no. 5, pp. 104-107, May 2000.

[18] A. Dimou, O. Nemethova, M. Rupp, "Scene Change Detection for H.264 Using Dynamic Threshold Techniques," Proc. of the 5th EURASIP Conference on Speech and Image Processing, Multimedia Communications and Service, Smolenice, Slovac Republic. July 2005, ISBN 80-227-2257-X.

[19] VQEG: "Final report from the Video Quality Experts Group on the validation of objective models of video quality assessment." 2000, available at http://www.vqeg.org/.

[20] W. J. Krzanowski, "Principles of Multivariate Analysis," Clarendon press, Oxford, 1988.

[21] M. Ries, C. Crespi, O.Nemethova and M. Rupp, "Content Based Video Quality Estimation for H.264/AVC Video Streaming," http://pub-et.tuwien.ac.at