Video Quality Estimation for Mobile 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
Email: {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 well-known methods are based on spatial features. This paper presents a new reference-free approach for quality estimation based on motion characteristics.**

The character of motion is determined by the amount and direction of the motion between two scene changes. In this paper, two methods are presented. The first method, presents the design of a quality metric based on content adaptive parameters, allowing for content dependent video quality estimation. The second method estimates video quality in two steps. Firstly, the content classification with character sensitive parameters is carried out. Finally, based on the content class, frame rate and bitrate, the video quality is estimated.

The performance of the proposed methods is evaluated and compared to the ANSI T1.801.03 metric. The results show that the motion-based approach provides powerful means of estimating the video quality experienced by users for low resolution video streaming services.

***Index Terms*— video quality, video streaming, perceived quality metric, MOS, QoS.**

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, video streaming for low bit and frame rates is allowed while preserving its perceptual quality. This is especially suitable for video applications in 3G wireless networks.

Mobile video streaming is characterized by low resolutions, and low bitrates. 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 or QVGA, 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 various streaming content and streaming application settings (resolution, frame and bit rate) as is demonstrated in [3], [4], [5], [6].

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 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 better reflect perceptual quality especially for low-rate videos. Most of these metrics were designed for broadband broadcasting video services and do not consider mobile video streaming scenarios.

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. 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 video quality simple enough to be calculated in real-time at the receiver side. We present new reference-free approaches for quality estimation based on motion characteristics. The first approach introduces a quality metric based on content adaptive parameters, allowing for content dependent video quality estimation. The second approach estimates video quality in two steps. Firstly, the content classification with character sensitive parameters is carried out. Finally, based on the content class, frame rate and bitrate, the video quality is estimated. Moreover, in this paper we provide a complex comparison of our recent models for video quality estimation [15], [16].

The paper is organized as follows: In Section 2 and 3 we describe a mobile video streaming scenario and a test setup for video quality evaluation, respectively. In Section 4 the process of motion feature extraction is explained. The results are presented and their performance evaluated and further processed in Section 5, where the focus is given on the video quality estimation. Section 6 contains conclusions and provides an outlook on future work.

II. MOBILE VIDEO STREAMING SCENARIO

Our mobile video streaming scenario is specified by the environment of usage, streamed content, and the screen size of the mobile terminal. Therefore, the mobile scenario is strictly different in comparison with classical TV broadcasting services or broadband IP-TV services. Furthermore, most of the mobile content is on demand. The mostly provided mobile streaming contents are news, soccer, cartoons, panorama for weather forecast, traffic news and music (see Figures 2, 3, 4, 5).
Our extensive survey shows systematic differences between MOS (Mean Opinion Score) results obtained by testing on UMTS terminals and PC screens [4]. According to these experiences, we perform our tests on UMTS mobile terminals. Due to this experience we did not follow ITU - T Recommendation [17] and in order to emulate real conditions of the UMTS service, all the sequences were displayed on a PDA VPA IV UMTS/WLAN (see Figure 1). The viewing distance from the phone is not fixed, but selected by the test person. We have noticed that the users are comfortable to take the UMTS terminal at a distance of 20-30 cm. The test was carried out in our video quality laboratory. Our video quality test design follows these experiences in order to better reflect real world scenarios.

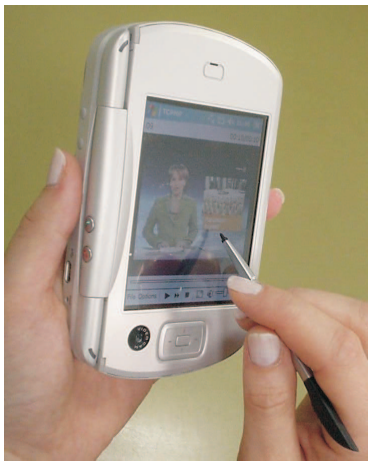

Figure 1. Test equipment: VPA IV UMTS/WLAN

For mobile video streaming we define the five most frequent contents with different impact on the user perception:

Figure 2. Snapshot of typical content class 1 (news)

1) Content class (CC1 = news): The first content class includes sequences with a small moving region of interest (face) on a static background. The movement in the region of interests (ROI) is mainly determined by eyes, mouth and face movements. The ROI covers up to approximately 15% of the screen surface.

Figure 3. Snapshot of typical content class 2 (soccer)

2) Content class (CC2 = soccer): This content class contains wide angle camera sequences with uniform camera movement (panning). The camera is tracking a small rapid moving objects (ball, players) on the uniformly colored (typically green) background.

Figure 4. Snapshot of typical content class 3 (cartoon)

3) Content class (CC3 = cartoon): In this content class object motion is dominant, the background is usually static. The global motion is almost not present due to its artificial origin of the movies (no camera). The movement object has no natural character.

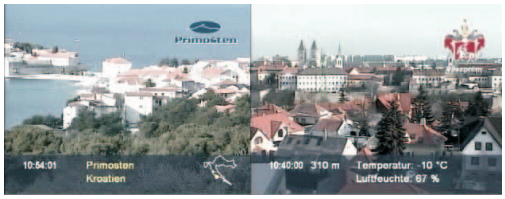

Figure 5. Snapshot of typical content class 4 (panorama)

4) Content class (CC4 = panorama): Global motion sequences taken with a wide angle panning camera. The camera movement is uniform and in a single direction.

5) Content class (CC5 = rest): The content class contains a lot of global and local motion or fast scene changes. Scenes shorter than three seconds are also associated to this content class. The content class covers scenes which do not fit any of the previous four classes.

III. TEST METHODOLOGY FOR VIDEO QUALITY EVALUATION

For the tests all sequences were encoded with the H.264/AVC baseline profile 1b. For subjective quality testing we used frame and bit rate combinations shown in Table III. In total there were 39 combinations.

TABLE I.
TESTED COMBINATIONS OF FRAME RATES (FR) AND BIT RATES(BR)

FR [fps]/BR [kbit/s]	24	50	56
5	CC1, CC3, CC4	CC5	CC1, CC2, CC3, CC4
7.5	CC1, CC3, CC4		CC1, CC2, CC3, CC4
10	CC1, CC3		CC1, CC2, CC3, CC4
15	CC1		CC1, CC2

FR [fps]/BR [kbit/s]	60	70	80	105
5				CC1
7.5	CC5	CC5		CC1, CC2, CC5
10		CC5	CC5	CC1, CC2, CC5
15			CC5	CC1, CC2, CC5

To obtain MOS values, we worked with 36 test persons for two different sets of test sequences. The first set was used for metric design and the second for evaluation of the metric performance. The training set test was carried out with 26 test persons and the evaluation test set was carried out with 10 test persons. The training and evaluation tests were collected from different sets of the five video classes. The chosen group of test persons ranged different ages (between 20 and 30), gender, education and experience with image processing.
The test method was Absolute Category Rating (ACR) as it better imitates the real world streaming scenario. Thus, the subjects did not have the original sequence as a reference, resulting in a higher variance. The sequences were presented in arbitrary order and the test environment followed ITU recommendation [17]. People evaluated the video quality after each sequence using a five grade MOS scale (1-bad, 5-excellent) in a prepared form.

A. Subjective quality test results

The obtained MOS data was scanned for unreliable and inconsistent results. Votes from one viewer to a certain sequence that differ two or more MOS grades from the first to the second run were considered unreliable and therefore rejected. In total, 12.3% of the results were rejected. This correction had negligible effect on the test global mean score. The 95% confidence intervals [17] were as well computed, assuming the votes follow a normal distribution. The MOS values obtained for all the test configurations ranged from 1.6 to 4.4. The distribution of the 95% confidence intervals for the MOS, can be used as a quality indicator of the collected data. The average size of the 95% confidence intervals is 0.27 on the 1-5 MOS scale. This indicates a good agreement between observers.

As can be seen from Figure 6, subjective video quality is strongly content dependent, especially for lower BR. For the ”news” sequence, the highest score is obtained by the configuration BR@FR=105@7.5 kbps@fps, closely followed by 105@10 kbps@fps and 56@10 kbps@fps. Very interesting is the fact that the viewer seems to notice no difference in quality between the combination 56@10 kbps@fps and 105@10 kbps@fps, which both receive very positive evaluations. The most dynamic sequence ”soccer” received the best evaluation at 105 kbps. An increasing frame rate has always a positive effect on the perceived quality, which is in contrast with other content types, specially to the ”news” case. In the ”soccer” sequence viewers prefer smoothness of motion rather than static quality.
The ”panorama” sequence receives better evaluation on lower FR. This indicates that the users give priority to the static quality in this case. In view of the ”cartoon” results, we can say that a sequence of these characteristics can be compressed at the very low data rate of 24 kbps, still obtaining a good perceived quality. At 56 kbps the static quality of the images is very good and does not worsen perceptibly with increasing frame rate. Therefore, at this data rate the quality perception of the viewers improves with FR and the configuration 56@10 kbps@fps receives the highest score a 4.4 MOS grade, which is even the absolute maximum score reached in the survey. The ”video clips” encoded at the highest rate 105 kbps have very good acceptance, but again we can observe better evaluation for 10 fps than for 15 fps.

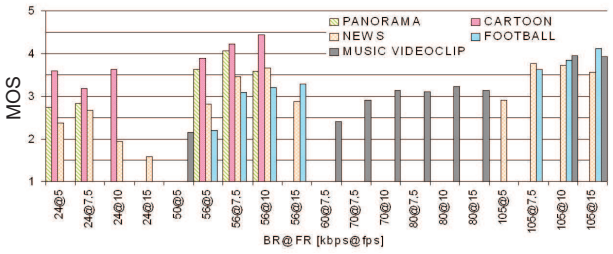

Figure 6. MOS for all the tested sequences (training set)

IV. VIDEO FEATURES EXTRACTION

In this section both our approaches for quality estimation are described. Both estimation methods use temporal segmentation before quality estimation. Furthermore, both methods are based on content/character sensitive parameters. The main difference between them is the estimation process alone.
The human visual perception of video content is determined by the character of the observed sequence. It is necessary to determine different content characters/classes or content adaptive parameters because the video content itself strongly influences the subjective quality. The character of a sequence can be described by the amount of the edges (spatial information) in the individual frames and by the type and direction of movement (temporal information). The data rate of the video sequence is shared by the number of frames per second. Higher frame

rates result in a lower amount of spatial information in individual frames and possibly in some compression artifacts.
In the literature the focus is given mainly on the spatial information [13], [14]. Such approaches come mainly from the quality estimation of still images [18], [19]. However, especially in small resolutions and after applying compression, not only the speed of movement (influencing at most the compression rate) but also the type of the movement plays 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.

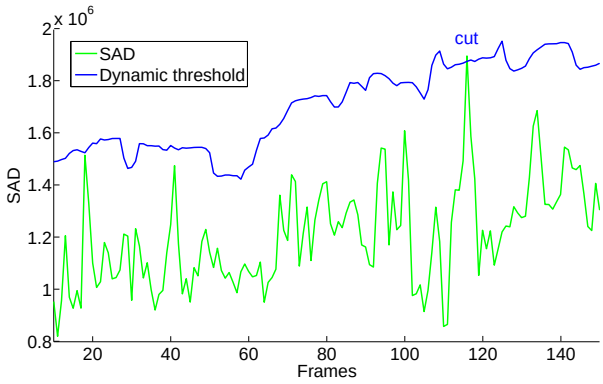

Figure 7. Performance of dynamic threshold function on a sequence with global rapid movement (car race).

A. Scene change detector

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 [20]. For our purpose the method was adopted 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 [20] (SAD) between two frames (n and $n+1$). Moreover, empirical mean m_n and standard deviation σ_n are computed for a sliding window $[n-N, n+N, N=10]$:

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

and

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

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

$$T_n = a \cdot m_n + b \cdot \sigma_n. \qquad (3)$$

The constants a, b were tuned in order to get the best performance for all content types. The constant a was set in order to avoid wrong scene 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 both precision and recall higher than 97%.

B. Motion vectors

As a first step we partitioned the current video frames in blocks of pixels, also known as the *target block*. The relative difference in the locations between the matching block and the target block is known as the *motion vector (MV)*. If the matching block is found at the same location as the target block then the difference is zero, and the motion vector is known as *zero MV*.
The difference between target and matching block increases (approximately linearly) with the size of the blocks and smaller blocks better describe the actual motion in the frame. On the other hand an increase of the objective accuracy does not always imply a better performance. We have observed that, if the blocks are selected too small, the resulting MVs do not reflect anymore the motion as it is perceived by a viewer. Due to the unavoidable presence of noise in video sequences, and the characteristics of the human visual system, it happens that movement is detected although a human observer does not see it. Such behavior is not suitable for our purpose. After several trials with videos of different character, we found a block size of 8×8 pixels to be a good trade-off for QVGA resolution sequences. The 320×240 pixels are divided into 30×40 blocks, which gives a total number of 1200 MVs per frame.
The second part of the process, and the most time and resource consuming one, is block matching. Each block in the current frame is compared to a certain search region in the past frame in order to find a matching block. This operation is performed only on the luminance component of the frame. A matching criterion has to be used to quantify the similarity between the target block and the candidate blocks. Because of its simplicity and good performance, we decided to use the sum of absolute differences (SAD), computed as the pixel wise sum of the absolute differences between the two blocks being compared:

$$\mathrm{SAD}_{n,m} = \sum_{i=1}^{N}\sum_{j=1}^{M} |B_n(i,j) - B_m(i,j)| \qquad (4)$$

where B_n and B_m are the two blocks of size $N \times M$, and i and j denote pixel coordinates. If more than one SAD minimum is detected, priority is given to the matching block the position of which is most similar to that of the target block, or equivalently, to the MV of smallest size.

C. Extraction of sequence motion and color parameters

Once we obtained MVs, the information about the motion (motion features) in the sequence has to be extracted. The static or dynamic character of a sequence is one of the main causes for the differences in perceived quality. We intended to perform a classification not only in terms of "static sequences" and "dynamic sequences", but also to investigate this aspect more in depth and determine typical levels of quantity of movement for every main content class. The overall amount of movement, or equivalently, the lack of movement in a frame, can be easily estimated from the proportion of blocks with zero vectors, that is, blocks that do not move from one frame to the other. Therefore, the average proportion of static blocks in a sequence of frames is very useful when it comes to distinguishing contents with typical different "levels" of overall movement.
The length of the MV indicates how far the block has moved from one frame to the next, and its angle tells us in which direction this movement occurred. Therefore, the mean size of the MVs in a frame or sequence of frames is an indicator of how fast the overall movement happens. On the other hand, knowing exactly in which direction the movement is taking place seems useless (redundant) for our purpose. Moreover, detecting a main direction of movement, that corresponds to big proportion of MVs pointing in the same direction, is a valuable information. Thus, it can be assumed that the analysis of the distribution of sizes and angles of the MVs can give substantial information about the character of the motion in the sequence. A set of statistical calculations on the MV was implemented in order to study their level of significance and find out which features can be used to identify perceptual content types and the video quality itself.

D. Sequence motion and color parameters for content classification

The content classification is based on the following statistical and resolution independent features of MVs within one shot (over all the frames of the analyzed sequence):

- **Zero MV ratio N_z:**
  Percentage of zero MVs in a frame. It is the proportion of the frame that does not change at all (or changes very slightly) between two consecutive frames. It usually corresponds to the background if the camera is static within one 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 n:**
  Proportion of mean size of the non-zero MVs within one frame normalized to the screen width, expressed in percentage. This parameter determines the amount of the global motion. 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.
- **Horizontalness of movement h:**
  We define horizontalness as the percentage of MVs pointing in horizontal direction. Horizontal MVs are from intervals $\langle -10; 10\rangle$ or $\langle 170; 190\rangle$ degrees.
- **Uniformity of movement d:**
  Percentage of MVs pointing in the dominant direction (the most frequent direction of MVs) in the frame. For this purpose, the granularity of the direction is 10 degrees.

In order to increase the accuracy of the content classifier, color features were considered. Color histograms provide additional information about the spatial sequence character because in different types of contents, the density and magnitude of colors differ as well. Soccer sequences for example contain a lot of varying green colors while cartoon sequences exhibit discrete saturated colors. This characteristic has important consequences to the compression and transmission artifacts. Therefore, we also use the following parameter:

- **Greenness g :**
  We define greenness as percentage of green pixels in a frame. For this purpose the RGB color space was down sampled to two bits per color component resulting in 64 colors. Five colors out of the 64 colors cover all variations of the green color.

E. Hypothesis testing and content classification

The content classification is based on the above defined parameters. Due to extensive set of objective parameters, the a statistical method was used for data analysis and content classification. This excludes content classifying based on threshold which is a limited and not accurate method for evaluating larger data sets.
We use a statistical method based on hypotheses testing. Each of the described content classes is determined by unique statistical features of motion and color parameters (see Figure 8). Due to their unique statistical features of well defined content classes it is not necessary to perform M-ary hypothesis testing and it is sufficient to formulate a null hypothesis (H0) for each content class based on these statistical features separately. The obtained empirical cumulative distribution functions (ECDF) from the typical set of sequences for each content class show substantial mutual differences (see Figure 8). From the next investigation it results that it is very difficult to determine single parametric distribution model representation from obtained model ECDF. For this purpose we were looking for hypotheses testing methods which allow for defining non-parametric, distribution free H0 hypotheses. For our hypothesis evaluation a method is needed capable of working with empirical (sample) distributions. For

this purpose the most suitable is the non-parametric and distribution free Kolmogorov-Smirnov (KS) test [21]. The KS test is used to determine whether two underlying probability distributions differ, or whether an underlying probability distribution differs from a hypothesized distribution, in either case based on finite samples. The two-sample KS test is one of the most useful and general non-parametric methods for comparing two samples, as it is sensitive to differences in both location and shape of the empirical cumulative distribution functions of the two samples.

From the typical set of sequences for each content class the ECDFs are obtained. The model ECDFs were derived from a set of 142 typical sequences. Each content class is described with five model ECDFs (zero MV ratio, mean MV size, uniformity of movement, horizontalness of movement, greenness), which correspond to their H0 hypothesis, respectively. Furthermore, it is necessary to find the maximal deviation ($D_{cc\ max}$) within one content class for all parameters (for each model ECDF). If the $F_n(x)$ is the model ECDF and $F(x)$ is the ECDF of the investigated sequence. D_n; is the maximal difference between $F_n(x)$ and $F(x)$:

$$D_n = \max_x \|F_n(x) - F(x)\| . \quad (5)$$

The content class estimation is based on a binary hypothesis test within the first four content classes. With the KS test the ECDFs of the investigated sequence and all model ECDFs of the first four content classes are compared. The KS test compares five ECDF (of defined MV or color parameters) of defined content classes specified by the H0 hypothesis with all five ECDFs of the investigated content. If the D_n obtained for the tested CC, is smaller than $D_{\text{CC max}}$ for each parameter, then the sequence matches this CC.

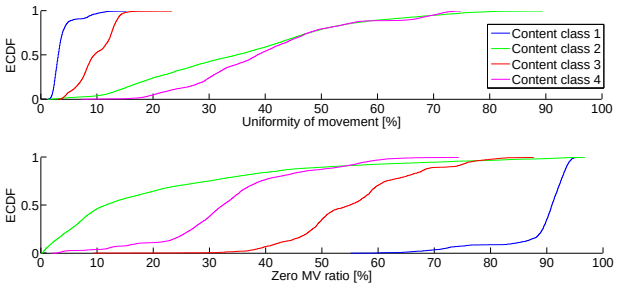

Figure 8. Model ECDF of zero MV ratio and uniformity of movement

If the ECDFs of the investigated sequence have not a fit with any of the first four content classes, the content classifier decides for the remaining content class number five. The classifier estimates the content at transmitter side from the original sequence.

The performance of the content classifier was evaluated with two parameters. **False detection** reflects the ratio of improper detection of a content class, in the case when investigated sequences belong to any **other** content class. **Good match** reflects the ratio of successful classification of investigated sequences, when investigated sequences belong to any of the first four classes. Note, in our sequences we had almost only cuts and no gradual changes. The scene change detector was sensitive on gradual shot boundaries (dissolve, fades or wipes). To evaluate the performance of the content classifier we used 786 sequences, 98% were classified correctly. The achieved precision of the content classifier is shown in Table II, what is a satisfying result for further quality estimation.

TABLE II.
THE EVALUATION RESULTS OF CONTENT CLASSIFIER

Content class	False detection [%]	Good match [%]
1	0	97
2	0	100
3	5.6	92
4	0	100

F. Sequence motion parameters for direct video quality estimation

For content classification we analyzed objective video parameters within one sequence (between two cuts). The difference is that for quality estimation we used data averaged over sequence and for content classification we process all sequence data to ECDF over all frames.

We focus on MV features, which make possible to detect rapid local movements or character of global movement. We investigated the 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 the influence of these parameters on the content.

For this purpose, we used a well known multivariate statistical method, the Principal Component Analysis (PCA) [22]. The PCA was carried out to verify further applicability of these characteristics for metric design. In our case the first two components proved to be sufficient for an adequate modeling of the variance of the data. The PCA results (see Figure 9) show influence of the chosen parameters (with the highest impact) on our data set for all content classes.

The following MV features and BR represent the motion characteristics:

- **Zero MV ratio within one shot Z:**
  Z equals the zero MV ratio N_z averaged over one shot .
- **Mean MV size within one shot V:**
  V equals the mean MV size averaged over one shot.

- **Ratio of MV deviation within one shot S:**
  Proportion of standard MV deviation 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 U:**
  This feature expresses proportion 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 expresses the 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 reflects the compression the compression gain in spatial and temporal domain. Moreover, the encoder performance is dependent on the motion characteristics. The BR reduction causes a loss of the spatial and temporal information, usually annoying for viewers.

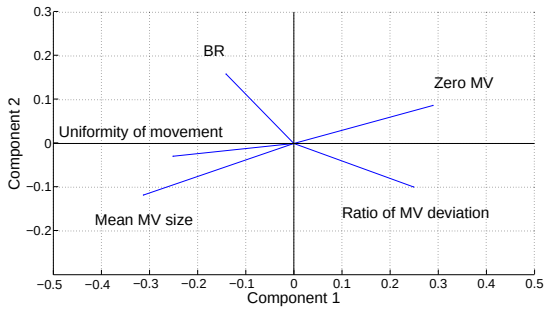

Figure 9. Visualization of PCA results for all content classes.

The perceptual quality reduction in spatial and temporal domain is very sensitive to the chosen motion features. This makes motion features very suitable for reference free quality estimation because higher compression does not necessarily reduce the subjective video quality (e. g. in static sequences).
The ANSI metric [11] consists of a linear combination of seven objective parameters based on spatial, temporal and chrominance properties of video streams. The biggest weight refers to spatial and chrominace component parameters. Furthermore, these parameters reflect both the video coding and network impairments. The ANSI defined "absolute temporal information feature" refers to the absolute amount of motion in a sequence but not to the character of motion as it is only based on the absolute differences of two consecutive frames.

V. Video Quality Estimation

We propose two methods for video quality estimation. The first is based on a set of reference free parameters. 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 decreases significantly the estimation accuracy. The proposed model reflects direct 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 (6) for all contents based on the defined motion parameters Z, S, V, U and BR:

TABLE III.
Coefficients of direct motion metric

Coeff.	Value
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 IV.
Coefficients of content based metric for all content classes (CC)

Coeff.	CC 1	CC 2	CC 3	CC 4	CC 5
A	4.0317	1.3033	4.3118	1.8094	1.0292
B	0	0.0157	0	0.0337	0.0290
C	−44.9873	0	−31.7755	0	0
D	0	0.0828	0.0604	0.0044	0
E	−0.5752	0	0	0	−1.6115

$$\widehat{\mathrm{MOS}}_{\mathrm{MV}} = \begin{array}{l} a + b \cdot \mathrm{BR} + c \cdot \mathrm{Z} + d \cdot \mathrm{S}^e + \\ + f \cdot \mathrm{V}^2 + g \cdot \ln(\mathrm{U}) + h \cdot \mathrm{S} \cdot \mathrm{V}. \end{array} \quad (6)$$

The metric coefficients (see Table III) 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 metric, we used a Pearson (linear) [23] correlation factor.
The second proposal is a content dependent low complexity metric based on two objective parameters (BR and FR) for each content class (7).

$$\widehat{\mathrm{MOS}}_{\mathrm{CC}} = f(\mathrm{BR}, \mathrm{FR}, \mathrm{CC}). \quad (7)$$

We proposed this common model (7) for each content class, each having a different parameter set A, B, C, D, E. Therefore, the model has linear and hyperbolic elements (8) and the coefficients vary substantially for the content classes. They can even have zero values. On the other hand, a rather good correlation was achieved with one offset and two non-zero coefficients (see Table IV).

$$\widehat{\mathrm{MOS}}_{\mathrm{CC}} = A + B \cdot \mathrm{BR} + \frac{C}{\mathrm{BR}} + D \cdot \mathrm{FR} + \frac{E}{\mathrm{FR}}. \quad (8)$$

The metric coefficients were obtained by a linear regression of the proposed model with our training set (MOS values averaged over two runs of all 26 subjective evaluations for particular test sequence). The model prediction performance on investigated content classes can be seen in Table V.

TABLE V.
CONTENT BASED METRIC PREDICTION PERFORMANCE BY CORRELATION ON EVALUATION SET

Metric/Content type	CC 1	CC 2	CC 3	CC 4	CC 5
Content based	0.9277	0.9747	0.9902	0.9030	0.9307
Direct motion based	0.8468	0.9812	0.9974	0.7140	0.9509

TABLE VI.
METRICS PREDICTION PERFORMANCE BY PEARSON CORRELATION

Metric	Pearson corr.
Direct motion based	0.8190
Content based	0.8303
ANSI	0.4173

A. Evaluation

The direct motion based metric [15] is a reference free estimator as well as the content based metric. However, the content based metric requires additional information on the content class [16]. The obtained prediction performance on the evaluation set (see Table VI and Figure 10) shows good agreement between MOS and estimated MOS results. Moreover, in Table V we can see a comparison of prediction performance of our models based on CCs. We can see that the content based metric has a good prediction performance on all CCs and is particularly better on CCs one and four (news and panorama). This reflects the problem of universal video quality estimators (without content classification) because the subjective perception of video is content dependent. The weak performance of the ANSI metric shows that this metric is not suitable for a mobile streaming scenario.
The usage of the mobile streaming services influence the subjective evaluation. Therefore, a universal metric like ANSI is not suitable for the estimation of mobile video quality. Only for higher MOS values which occur at high bitrates ($\geq$ 90 kbps) the ANSI metric performs comparable to our proposed metrics (see Figure 10).
Note that the depicted values (see Figure 10) of our results appear a bit shifted upwards. This is due to the limited size of the evaluation set. The size of our training set is sufficient and our model performance on it appears to be unbiased.

B. Metric applicability

Our proposals for quality estimation are trade-offs between applicability, processing demands and prediction accuracy. The first proposal is more complex but allows us to divide content classification and quality estimation (see Figure 11). The suitable solution is to perform the content class classification at the streaming server and stream content information with video. Finally, the quality estimation is performed at user equipment. This allows us to estimate quality at the receiver with extremely low complexity. The next approach allows for a full-reference free estimation at the sender and receiver side. Currently, this proposal is more suitable for streaming servers due to limitations in processing power at the user devices.

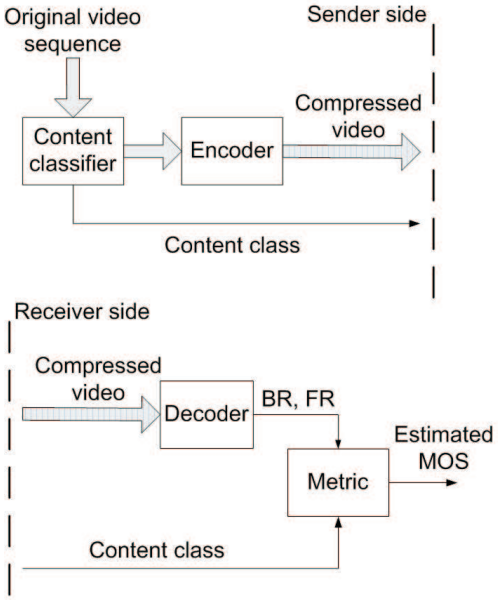

Figure 11. Content based video quality estimator design

In comparison to the well-known ANSI metric our proposals are less complex and more accurate, although the ANSI metric was not designed for video streaming services.

VI. CONCLUSIONS

The scope of this work was to estimate video quality for the most frequent content types in mobile video streaming scenarios. First, it was necessary to investigate and define mobile scenarios and a test methodology in order to achieve the best emulation of ”the real world scenario”. Furthermore, we were able to define content adaptive motion parameters which are based on MV features.
Finally, we propose two reference-free estimation methods. The first method estimates video quality in two steps — the content class is estimated from the original video sequence at the sender side, and then the quality metric is calculated at the receiver with almost zero complexity. The second, direct motion method is suitable for stand alone estimation at the receiver side. Furthermore, the direct motion proposal has a slightly worse estimation performance but allows full reference-free estimation for all content classes. The performance of both introduced video quality metrics shows good agreement between estimated MOS and the evaluation set. Moreover, both methods outperform the well-known ANSI metric.

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.

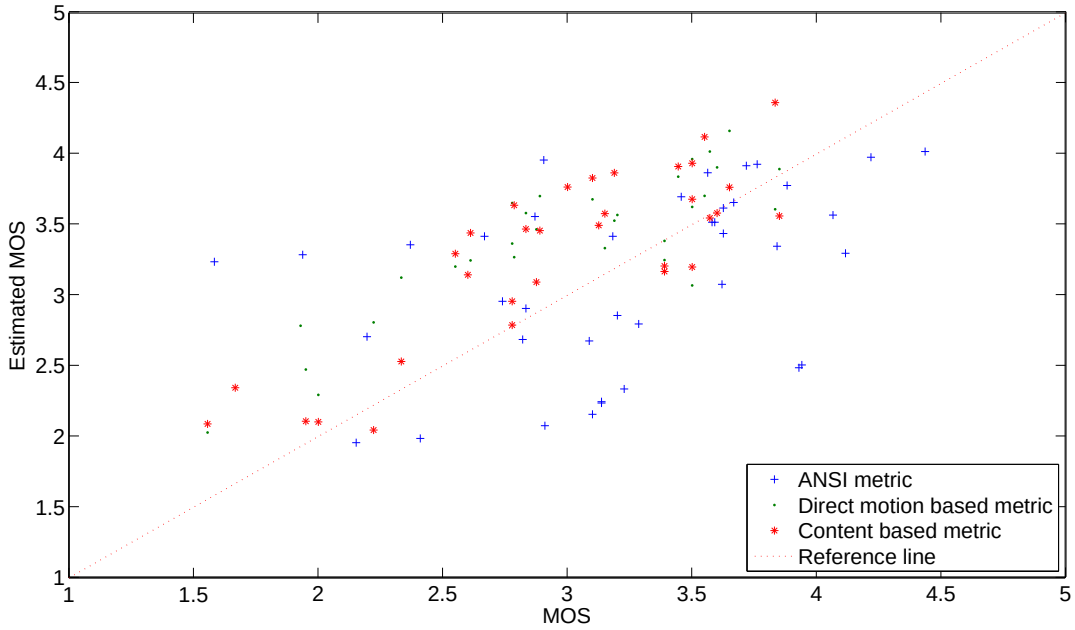

Figure 10. Estimated vs. subjective MOS results

REFERENCES

[1] 3GPP TS 26.234 V6.11.0:"End-to-end transparent streaming service; Protocols and codecs," Jun. 2007.

[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, Dec. 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, Nov. 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, 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, Jul. 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 2005 Systems Communications (ICW05), pp. 178-184, Montreal, Canada, Aug. 2005.

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

[15] M. Ries, O.Nemethova and M. Rupp, "Motion Based Video Quality Estimation for H.264/AVC Video Streaming," Proc. of the International Symposium on Wireless Pervasive Computing 2006, San Juan, Puerto Rico, Feb. 2007.

[16] M. Ries, C. Crespi, O.Nemethova and M. Rupp, "Content Based Video Quality Estimation for H.264/AVC Video Streaming," Proc. of the IEEE Wireless Communication & Networking Conference, Hong Kong, China, Mar. 2007.

[17] ITU-T Recommendation P.910, "Subjective video quality assessment methods for multimedia applications," Sep. 999.

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

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

[20] A. Dimou, O. Nemethova, M. Rupp, "Scene Change Detection for H.264 Using Dynamic Threshold Techniques," in Proc. of the 5th EURASIP Conference on Speech and Image Processing, Multimedia Communications and Service, Smolenice, Slovac Republic. Jul. 2005.

[21] K. Bosch, "Statistik-Taschenbuch," Oldenbourg Wissensch. Vlg, Munich, 1998.

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

[23] 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/.

Michal Ries is currently a research assistant and working towards his Dr.techn. (Ph.D.) degree at Vienna University of Technology, Institute of Communications and Radio-Frequency Engineering. He received his B.S. and M.S. degrees in 2002 and 2004 at the Slovak University of Technology, Faculty of Electrical Engineering and Information Technology in Bratislava. Before he joined TU Vienna he was working for Siemens PSE as system engineer.
His research interests include perceptual video and audiovisual quality evaluation, video and audiovisual metric design, monitoring of QoS in wireless networks, video streaming in wireless network optimisation.

Olivia Nemethova received her B.S. and M.S. degree from Slovak University of Technology in Bratislava in 1999 and 2001 respectively, both on Informatics and Telecommunications. She received her PhD on Electrical Engineering from Vienna University of Technology in 2007. From 2001 until 2003 she was with Siemens as a system engineer. She worked on UMTS standardization within 3GPP TSG RAN2 as a Siemens delegate. In parallel she worked within an International Property Rights management team responsible for evaluation of IPRs regarding radio access networks. In 2003 she joined the Institute of Communications and Radio- Frequency Engineering at Vienna University of Technology as a research and teaching assistant. Her current research interests include error resilient transmission of multimedia over wireless networks, video processing and mobile communications.

Markus Rupp received his Dipl.-Ing. degree in 1988 at the University of Saarbruecken, Germany and his Dr.-Ing. degree in 1993 at the Technische Universitaet Darmstadt, Germany, where he worked with Eberhardt Haensler on designing new algorithms for acoustical and electrical echo compensation. From November 1993 until July 1995 he had a postdoctoral position at the University of Santa Barbara, California with Sanjit Mitra where he worked with Ali H. Sayed on a robustness description of adaptive filters with impacts on neural networks and active noise control. From October 1995 until August 2001 he was a member of the Technical Staff in the Wireless Technology Research Department of Bell-Labs at Crawford Hill, NJ, where he was working on various topics related to adaptive equalization and rapid implementation for IS-136, 802.11 and UMTS. He is presently a full professor for Digital Signal Processing in Mobile Communications at the Technical University of Vienna. He was associate editor of IEEE Transactions on Signal Processing from 2002-2005, is currently associate editor of JASP EURASIP Journal of Advanced Signal Processing, JES EURASIP Journal on Embedded Systems, Research Letters in Signal Processing, Research Letters in Communications, and is elected AdCom member of EURASIP. He authored and co-authored more than 250 papers and patents on adaptive filtering, wireless communications and rapid prototyping as well as automatic design methods.