PERFORMANCE EVALUATION OF MOBILE VIDEO QUALITY ESTIMATORS

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
web: www.nt.tuwien.ac.at

ABSTRACT

Provisioning of mobile video streaming is hitting toward to limitations in channel quality and capacity as well as in terminal processing power. These known limitations, network settings, and video content influence the end user quality. In this article we investigate the estimation of perceived video quality for mobile streaming scenarios. Firstly, we analyze streaming content and usage scenarios. Secondly, we define objective video parameters which reflect the sequence motion character and its content. Finally, video quality estimation methods based on these parameters are developed and compared with common methods.
The presented results show that the proposed approach provides powerful solutions for automatic subjective video quality estimation.

1. INTRODUCTION

Provisioning of mobile multimedia services is becoming more important, due to market and terminal development in recent years. The most challenging quality issue for provisioning of multimedia services is providing video streaming services in a required level of customer satisfaction. The required level of customer satisfaction is achieved with suitable combinations of codec and network settings for streamed content. This reflects a customer centric point of view on providing of video streaming services. Moreover, it requires an automatic estimation of perceived video quality, in order to maximize subjective performance of mobile video streaming.
Mobile video streaming is characterized by low resolutions and low bit rates. The bit rates are limited by the capacity of UMTS (Universal Mobile Telecommunications System) radio bearer and restricted processing power of mobile terminals. These properties exclude provisioning of 3G video streaming with bitrates (BR) higher than 200kbps. The commonly used resolutions are *Quarter Common Intermediate Format* (QCIF,176×144 pixels) for cell phones, *Common Intermediate Format* (CIF, 352×288 pixels) and *Standard Interchange Format* (SIF, 320×240 pixels) for data-cards and palmtops (PDA). The codec mandatory for UMTS video streaming is H.263. The video streaming encoded with H.263 contains a lot of video coding artifacts due to high compression ratios common in mobile streaming. This results to significant video quality reduction. Nevertheless, H.263 is widely used because it is license-free. UMTS optional codecs are MPEG4 and since release 6 [1] a baseline profile of the H.264/AVC codec [2] is supported. The most suitable video coding standard for mobile streaming is H.264/AVC. Thanks to its significant improvement in video compression gain the newest video coding standard H.264/AVC allows to provide video streaming for low bit and frame rates while preserving perceptual quality.
Our recent achievements [3], [4] show that the perceived video quality is influenced by mobile usage scenario and content type. Moreover, we can assume that individual objective video parameters are weakly correlated with subjective quality [5]. The proposed metrics of the last years can be subdivided into two main groups: human vision model based video metrics [6], [7], [8] and metrics based only on the objective video parameters [9], [10], [11]. The complexity of these methods is quite high and significant computational power is necessary to calculate them. These metrics are designed for broadband broadcasting video services and do not consider mobile video streaming scenarios.
Due to content dependence of subjective video quality [5], [12] it is highly meaningful to use sequence character sensible objective parameters [3] or additive content classification [4].
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. Furthermore, we investigated different estimation methods based on complex analytical models [3], low-complexity analytical models with content classifications [4] and novel ensemble based estimation systems. The last estimation method shows that ensemble based systems are more beneficial than their single classifier counterparts. Moreover, we discuss the suitability of these estimation methods for automatic video quality in different usage scenarios.
The paper is organized as follows: in Section 2, our video streaming usage scenario and the sequences selected for evaluation are described as well as the setup of our survey we performed to obtain MOS values. In Section 3, the extraction process of character sensible objective parameters is explained. In Section 4, estimation methods and their performance are discussed. Section 5 contains conclusions and some final remarks.

2. 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 or traffic news and video clip.
Our extensive survey shows systematic differences between MOS (mean opinion score) results obtained by testing on

UMTS terminals and PC screens. According to these experiences, we perform our tests on UMTS mobile terminals. Due to this experience we did not follow ITU-T Recommendation [13] 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 and the tests were conducted in our video quality laboratory.
Our video quality test design follows these experiences in order to reflect real world scenario.

Figure 1: Test equipment: VPA IV UMTS/WLAN

2.1 The test setup for video quality evaluation

For the tests we selected two sets of five video sequences each having ten-second duration and SIF resolution. Screenshots of some of these sequences are depicted in Figure 2. All sequences were encoded with an H.264/AVC baseline profile 1b. For subjective quality testing we used frame and bit rate combinations shown in Table 1. In total there were 36 combinations.
We subdivided the sequences to theirs contents into five classes.

Figure 2: Snapshots of training and test sequences

In the ”news” or content class number 1 (CC #1) sequences, a moderator is reading news only by moving her lips and eyes. The ”news” sequences include parts with a small moving region of interest (face) on static background.
The ”soccer” or CC #2 sequences contain wide angle camera sequences with uniform camera movement (panning). The camera is tracking a small rapid moving object (ball) on the uniformly colored (typically green) background.
In ”cartoon” or CC #3 sequences the object motion is dominant, the background is usually static. 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” or CC #4 sequences contain global motion sequences taken with a wide angle panning camera. The camera movement is uniform and in one direction.
The last investigated sequence videoclip CC #5 either contains a lot of global and local motion or fast scene changes.

FR [fps]/BR [kbit/s]	24	50	56
5	CC #1, CC #4	CC #5	CC #1, CC #4
7.5	CC #1, CC #4		CC #1, CC #4
10	CC #1, CC #4		CC #1, CC #4
15	CC #1		CC #1

FR [fps]/BR [kbit/s]	60	70	80	105
5				CC #1
7.5	CC #5	CC #5		CC #1, CC #2, CC #5
10		CC #5	CC #5	CC #1, CC #2, CC #5
15			CC #5	CC #1, CC #2, CC #5

Table 1: Tested combinations of frame rates (FR) and bit rates(BR)

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 test set 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 five video sequences. 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 had not the original sequence as a reference, resulting in a higher variance. People evaluated the video quality after each sequence using a five grade MOS scale (1-bad, 2-poor, 3-fair, 4-good, 5-excellent) in a prepared form.

2.2 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 [13] 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, shown in Figure 3, 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 4, subjective video quality is strongly content dependent, especially for lower BR. For the ”news” sequence highest score is obtained by the configuration 105@7.5, closely followed by 105@10, 56@10

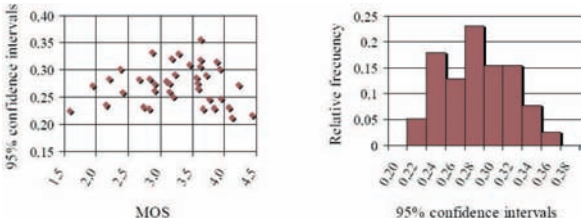

Figure 3: Left, confidence interval size vs. MOS. Right, distribution of 95% confidence intervals.

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 and 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 hat the users give in this case priority to the static quality. 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, 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 viewers quality perception improves with FR and the configurations 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 at 15 fps.

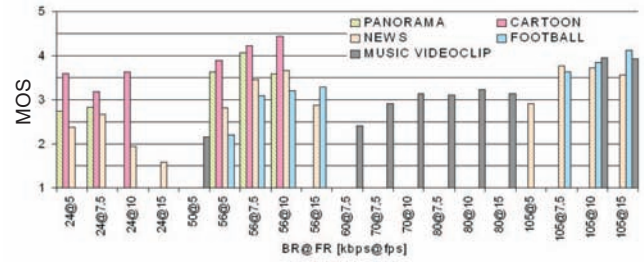

Figure 4: MOS for all the tested sequences (training set)

3. OBJECTIVE VIDEO QUALITY PARAMETERS

As we can observe from the obtained MOS data, the human perception is strongly influenced by the character of the sequence. Especially in small resolutions and after compression, not only speed of movement (influencing directly the compression gain) 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. The motion features can be used directly as an input into the estimation formulas or indirectly by determining limited number of content classes and using distinctive MOS estimation metrics for each content class. Both possibilities we investigated in [3] and [4], respectively.
The investigated motion features concentrate on the motion vector statistics, including the size distribution and the directional features of the motion vectors (MV) within one sequence of frames between two cuts. Zero MVs allow for estimating the size of the still regions in the video pictures. That, in turn, allows to analyze MV features for the regions with movement separately. This particular MV features makes it possible for distinguishing between rapid local movements and global movement. For content classification, the motion features are calculated for each frame and statistically processed for further hypotheses testing.

4. VIDEO QUALITY ESTIMATION

The perceived video quality estimation must be based on character sensitive motion features within one shot (scene), because a video stream can consist of more than one different shot having different content. Therefore, we estimate video quality within one scene. For our purpose we extend algorithm for temporal segmentation based on a dynamic threshold [14]. For *content class based* video quality estimation, the content class is estimated firstly as is explained in [4]. The automatic content classification enables video quality estimation within one content class. The classification based on hypothesis testing is a universal statistical method for content classification, which provides almost unlimited opportunities for the definition of new content classes. The subjective quality for a certain content class is then estimated as a function of BR and FR. The content class based metric is a reference-free estimator if the content class is known (e.g. signalized).
The *motion based* quality estimation is a fully reference-free method which allows for sensible quality estimation for the most diverse content classes. Our first approach was to use one single analytic model [3] to reduce the estimation complexity. In order to obtain higher accuracy of our estimation, we investigated an estimation using ensemble based systems. Ensemble based systems combine the outputs of several classifiers (estimators) by averaging in order to reduce the risk of an unfortunate selection of a poorly performing classifier.

4.1 Ensemble based video quality estimation

The very first idea to use more than one classifier for estimation comes from the neural network community [15]. In the last decade research in this field has expanded in strategies [16] for generating individual classifiers, and/or the strategy employed for combining the classifiers.
Our approach is to train a defined ensemble of models with a set of four motion sensitive objective parameters and BR [3]. We build our ensemble of different model classes, to improve the performance in regression problems. The theoretical background [17] of our approach is that an ensemble of heterogeneous models usually leads to a reduction of the ensemble variance because the cross terms in the variance contribution have a higher ambiguity. Furthermore, we demonstrate that an ensemble of models has a better performance than a single model. We consider a dataset with input values (motion sensitive parameters and BR) $\mathbf{x}$ and output value (MOS) y with a functional relationship, where r is an estimation error:

$$y = f(\mathbf{x}) + r. \quad (1)$$

The weighted average $\bar{f}(\mathbf{x})$ of the ensemble of models is defined as follows:

$$\bar{f}(\mathbf{x}) = \sum_{k=1}^{K} w_k f_k(\mathbf{x}), \tag{2}$$

where $f_k(\mathbf{x})$ denotes the k-th individual model and the weights w_k sum to one ($\sum_k w_k = 1$). The generalization (squared) error $e(\mathbf{x})$ of the ensemble:

$$e(\mathbf{x}) = (y(\mathbf{x}) - \bar{f}(\mathbf{x}))^2. \tag{3}$$

According to [17], the error can be decomposed as follows:

$$e(\mathbf{x}) = \bar{e}(\mathbf{x}) - \bar{a}(\mathbf{x}). \tag{4}$$

This assumption allows us to neglect the mixed terms of following equation, where the average error $\bar{e}(\mathbf{x})$ of the individual model is:

$$\bar{e}(\mathbf{x}) = \sum_{k=1}^{K} w_k (y(\mathbf{x}) - f_k(\mathbf{x}))^2, \tag{5}$$

and the average ambiguity $\bar{a}(\mathbf{x})$ of the ensemble is:

$$\bar{a}(\mathbf{x}) = \sum_{k=1}^{K} w_k (f_k(\mathbf{x}) - \bar{f}(\mathbf{x}))^2. \tag{6}$$

According to (5) and (6) we can make the following assumptions:

- The ensemble generalization error $e(\mathbf{x})$ is always smaller than the expected error of the individual models $\bar{e}(\mathbf{x})$,
- An ensemble should consist of well trained but diverse models in order to increase the ensemble ambiguity.

These assumptions were applied to an ensemble of universal models. In order to estimate the generalization error and to select models for the final ensemble we used a cross-validation scheme for model training [18]. These algorithms increase ambiguity and thus improve generalization of a trained model. Furthermore, we obtain an unbiased estimator of the ensemble generalization error.
The cross-validation works as follows:

- Our data set is divided in two subsets and the models are trained on the first set.
- The models are evaluated on the second set, the model with the best performance becomes ensemble member.
- The data set is divided with light overlapping with previous subsets into two new subsets and the models are trained on the first set.
- The cross-validation continues until the ensemble has a desired size. The best trade-off between ensemble complexity and performance was achieved for ensemble of six estimators.

4.1.1 The diversity of the ensemble set

The final step in the design of an ensemble based system is to find a suitable combination of models. Due to outliers and overlapping in data distribution of our data set, it is impossible to propose a single estimator with perfect generalization performance. Therefore, we exploit an ensemble of many classifiers and combine their outputs such that the combination improves upon the performance of a single classifier. Moreover, we are looking for classifiers with significantly different decision boundaries from the rest of the ensemble set. This property of an ensemble set is called diversity. The above mentioned cross-validation introduces model-diversity, the training on slightly different data sets leads to different estimators (classifiers). Additionally, we increase diversity by using two independent models. Furthermore in cross validation we automatically exclude classifiers with worse correlation than 50% on the second set.
As the first estimation model we chose a simple, nonparametric method the k-nearest neighbour rule (kNN) with adaptive metric [18]. This method is very flexible and does not require any preprocessing of the training data. The kNN decision rule assigns to an unclassified sample point the classification of the nearest sample point of a set of previous classified points. We used a locally adaptive form of k-nearest neighbor classification. The k is selected by cross validation.
As the second method an artificial neural network (ANN) is used. We designed the network with three layers – input, one hidden and output layers with five objective parameters as an input and estimated MOS as output. Each ANN has 90 neurons in the hidden layer. As a learning method we used improved resilient propagation (IRPROP+) with back propagation [19]. IRPROP+ is a fast and accurate learning method in solving estimation tasks for our data set. Finally the ensemble consist of two estimation models kNN and ANN and six estimators, three kNN and three ANN.

Metric	Pearson corr. [%]
Ensemble based	85,54
Motion based	80,25
Cont. cl. based	81,93
ANSI [9]	41,73

Table 2: Performance of the MOS estimator for all content classes.

Metric	CC 1	CC 2	CC 3	CC 4	CC 5
Ensemble based	0.93	0.97	0.77	0.91	0.97
Motion based	0.77	0.97	0.86	0.80	0.94
Content class based	0.93	0.90	0.76	0.90	0.93
ANSI [9]	0.63	0.85	0.95	0.93	0.97

Table 3: Pearson correlation factor of the MOS estimators for particular content classes.

4.2 Performance of the video quality estimators

To validate the performance of our proposed metric, we used the Pearson (linear) correlation factor [20]. In order to provide a detailed comparison, we compare performance of our ensemble based estimator with content class based [4] and motion based [3] estimators and the ANSI metric [9] on our evaluation set. The depicted results for Pearson correlation factor in Table 2 reflect the goodness of fit (see Figure 5) with the independent evaluation set for all content types together. This correlation method only assumes a monotonic relationship between the two quantities. A virtue of this form of correlation is that it does not require the assumption of any particular functional form in the relationship between data and predictions. The results in Table 2 clearly show good monotonicity between obtained and estimated values for all

tested metrics.
In addition, we investigated the goodness of our fit on different content classes (see Table 3). The best performance over all content classes is provided by ensemble and content class based metrics. A fair performance was obtained by the motion based metric and very unbalanced performance by the ANSI metric. The ensemble based metric has performance similar to the content class based metric. The content classification can be understood as an art of pre-estimation in order to obtain a more homogeneous set of results within one content class, which allows for more accurate quality estimation. This effect was achieved by introducing cross-validation in ensemble based metrics. The motion based metric suffers from weakness of single model estimation. The fit for 'news' and 'cartoon' sequences is relatively lower in comparison to the other content classes. A closer look on the ANSI metric performance shows that ANSI metric has good fit with CCs #3, 4, 5 and poor performance on the rest of CCs. Moreover, the ANSI metric requires the knowledge of a reference video (original) and is the most complex estimator.

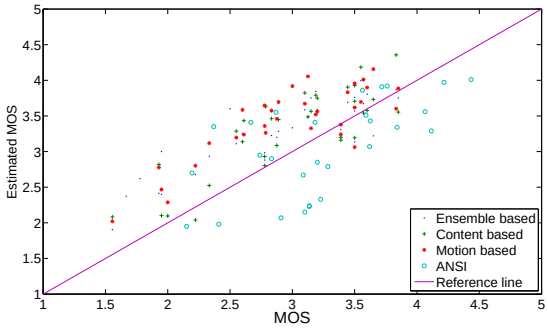

Figure 5: Estimated vs. subjective MOS results

5. CONCLUSIONS

In this paper we proposed an ensemble based perceptual quality metric for the most frequent content types for mobile video streaming services and investigated its performance. Furthermore, we compared the results with the other state of the art metrics. The investigated metrics differ in complexity and applicability. The ensemble based metric and motion based metric are fully reference-free estimators. The ensemble based metric is more complex but the most accurate method. The ensemble based and content class based methods have similar complexity but are still less complex than the reference ANSI metric. The universal ANSI metric is not suitable for video quality estimation in the chosen scenarios due to its complexity and different perception of the mobile video streaming services.

6. 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:"Transparent end-to-end Packet-switched Streaming Service (PSS); 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, "Motion Based Reference-Free Quality Estimation for H.264/AVC Video Streaming," Proc. of the ISWPC'07, San Juan, PR, USA, Feb. 2007.

[4] M. Ries, C. Crespi, O. Nemethova, M. Rupp, "Content Based Video Quality Estimation for H.264/AVC Video Streaming," Proc. of the IEEE WCNC'07, Hong Kong, Mar. 2007.

[5] 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.

[6] 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.

[7] 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.

[8] S. Winkler, "Digital Video Quality," John Wiley & Sons, Chichester, 2005.

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

[10] M.H. Pinson, S. Wolf, "A new standardized method for objectively measuring video quality," IEEE Trans. on Broadcasting, vol. 50, no. 3, pp. 312–322, Sep. 2004.

[11] 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.

[12] 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.

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

[14] A. Dimou, O. Nemethova, M. Rupp, "Scene Change Detection for H.264 Using Dynamic Threshold Techniques," Proc. of the 5th EURASIP Conf. on Speech and Image Proc., Multimedia Comm. and Services, Smolenice, Slovakia. July 2005.

[15] B.V. Dasarathy, B.V. Sheela, "Composite classifier system design: Concepts and methodology," Proc. of the IEEE, vol. 67, no. 5, pp. 708-713, 1979.

[16] L.I. Kuncheva, "Combining Pattern Classifiers, Methods and Algorithms," New York, Wiley Interscience, 2005.

[17] Krogh, Vedelsby: "Neural Network Ensembles, Cross Validation and Active Learning," Advances in Neural Information Processing Systems 7, MIT Press, 1995.

[18] Hastie, Tibshirani, Friedman: The Elements of Statistical Learning, Springer, 2001.

[19] C. Igel, M. Hsken, "Improving the Rprop learning algorithm," in Proc. of the 2nd Int. Symp. on Neural Computation, (pp. 115-121), Berlin, ICSC Academic Press, 2000.

[20] 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/.