Reference-Free Video Quality Metric for Mobile Streaming Applications

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*— Mobile multimedia streaming applications are becoming more and more popular although the perceptual video quality for such low bit rates, frame rates and resolutions is limited. Depending on the content character of a video sequence, the compression settings maximizing the subjective perceptual quality also differ. Aim of this work is to find a low-complexity reference-free estimation of visual perceptual quality, based on a combination of a possibly small set of the most important objective parameters - compression settings and content features. To achieve this, various objective parameters are mapped on mean opinion score (MOS), obtained by an extensive survey. Finally, an estimation is chosen having the best correlation with the measured data set. The so obtained quality metric does not require any knowledge about the original sequence. Its correlation with the data set is as good as if more complex human vision based estimation was applied.**

I. Introduction

Mobile video streaming became reality in emerging 3rd generation networks. In UMTS services, it is essential to provide required level of customer satisfaction, given in case of video by the perceived video stream quality. It is therefore important to choose the compression parameters as well as the network settings so that they optimize the end-user quality. Hence, we are looking for an objective measure of the video quality simple enough to be calculated in real-time. We need to distinguish different content character because they strongly influence the subjective quality [1]. Goal of our research is to estimate the video quality of mobile video streaming at the user-level (perceptual quality of service) and to find most suitable codec settings for the most frequent content types. Mobile video streaming is characterized by the low resolutions that can be used - mostly 144×176 QCIF for telephones or 288×352 CIF for data-cards and palmtops. In UMTS the bearers with 64-384 kbit/s are used for multimedia (audio and video) streaming as the capacity of 2 Mbit/s has to be shared by all the users in a single cell. Mobile terminals also have limited complexity and power, so the decoding of higher rate videos becomes quite a challenging task.

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 [2], [3] and metrics based only on the objective video parameters [4], [5]. The complexity of these methods is quite high and significant computational power is necessary to calculate them. For smaller resolutions, it is possible to find simpler estimates achieving the same performance. This is the aim of our work. Moreover, we are looking at the measures that would 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.

The paper is organized as follows: In Section 2 the sequences selected for evaluation are described as well as the setup of survey, we performed to obtain the MOS values. Section 3 describes objective video parameters and the data analysis. The results are introduced and further interpreted in Section 4. Focus is given on the metric design. Section 5 contains conclusions and some final remarks.

II. The test setup for video quality evaluation

For the tests we selected five video sequences each having ten-second duration and QCIF resolution. Screenshots of these sequences are depicted in Figure 1.

Fig. 1. Screenshots of the video test sequences used in the survey: ”akiyo”, ”foreman”, ”soccer”, ”panorama” and ”traffic”.

Two of them (”akiyo”, ”foreman”) are well-known professional test sequences obtained by a static camera. In the ”akiyo” sequence a female moderator is reading news only by moving her lips and eyes. The ”akiyo” sequence represents the news scenario. The ”foreman” sequence contains a monologue of a man moving his head dynamically and at the end

of the sequence there is a contiguous scene change. The ”foreman” sequence is typical scenario for video call. ”Soccer” and ”panorama” are both sequences with permanent camera movement. The ”Soccer” is a professional wide angle sequence; the entire picture is moving uniformly. Additionaly the players and ball are moving in a fast way. Then ”panorama” is a non-professional sequence, containing smooth and relatively slow movement of the whole scene. This is a typical scenario for weather cameras, wide angle survivalence and for tourists guides. The ”traffic” sequence is obtained by a static traffic camera. The camera is static and slowly moving cars can be observed. Each of the tested sequences represent typical content offered by network providers nawadays.

All sequences were encoded with H.263 profile 3 and level 10. For subjective quality testing we used combinations of bit rates and frame rates shown in Table I. In total, there were 60 encoded test sequences but we excluded six combinations where the resulting video quality was clearly insufficient.

Frame Rate [frames/s]	Bit Rate [kbit/s]
5	18
5	44
5	80
7.5	18
7.5	44
7.5	80
10	18
10	44
10	80
15	18
15	44
15	80

TABLE I
TESTED COMBINATIONS OF FRAME RATES AND BIT RATES.

To obtain MOS, we worked with 38 paid test persons. The chosen group ranged different ages (between 17 and 30), sex, education and experience with image processing.
The tests were performed according to the ITU-T Recommendation [6], using absolute category rating (ACR) method as it better imitates the real world streaming scenario. Thus, the subjects had not the original sequence as a reference, so their evaluation suffers from higher variance. People evaluated the video quality using a five grade MOS scale (1-bad, 2-poor, 3-fair, 4-good, 5-excellent). According to our experiences with previous psycho-visual experiments [1], [8], [10] the subjective results are slightly different if they are displayed on UMTS handset or PC monitor. To emulate real conditions of the UMTS service, all the sequences were displayed on a UMTS handset Sony Ericsson Z1010. The viewing distance from the phone was not fixed, but selected by the test person. We have noticed that all subjects were comfortable to take the phone at a distance of 20-30 cm. At the beginning of the test session, three training sequences were presented to the test persons. Test sequences were presented in an arbitrary order, with the additional condition that the same sequence (even differently degraded) did not appear in succession. Three runs of each test were taken. In order to avoid the learning effect we made a break of half an hour between the first and the second run, and a break of two weeks between the second and the third run. However, there were no really noticeable differences between the first two runs and the third run, performed two weeks after. In the further processing of our results we have rejected the sequences which were evaluated with individual standard deviation higher than one. Following this rule, we excluded 2,23% of the tests results.

III. CHOICE OF OBJECTIVE PARAMETERS SET

By the term ”objective parameters” we understand both - the compression parameters and the content characteristics. In total, we investigated eight objective video parameters with various computational complexity. We also looked at the five of the objective video parameters recommended in [4], that match the video quality for our scenario [8]. The first two of them are $sigain$ and $siloss$ measuring the gain and the loss in the amount of spatial activity, respectively. If the codec operates through an edge sharpening or enhancement, a gain in the spatial activity is obtained, that is an improvement in the video quality of the image. On the other hand, when a blurring effect is present in an image, it leads to a loss in the spatial activity. The other two parameters, $hvgain$ and $hvloss$ measure the changes in the orientation of the spatial activity. In particular, $hvloss$ reveals if horizontal and vertical edges suffer of more blurring than diagonal edges. The parameter $hvgain$ reveals if erroneous horizontal and vertical edges are introduced in the form of blocking or tiling distortions. These parameters are calculated over the space-time (S-T) regions of original and degraded frames. The S-T regions are described by the number of pixels horizontally, vertically and by the time duration of region. In our case one S-T region corresponds to 8×8 pixels over five frames. The complexity to calculate these parameters is rather high. Please, note that these parameters require the knowledge of the original sequence.

The fifth ANSI parameter we looked at, is a reference-free measure of overall spatial information, denoted as f_{SI13} since images were preprocessed using the 13×13 Sobel filter masks. It is calculated as the standard deviation over an S-T region of $R(i,j,t)$ samples, i and j being the coordinates within the picture displayed in time t. The result is clipped at the perceptibility threshold P [4]:

$$\mathrm{f_{SI13}} = \{\mathrm{std}_{space}[R(i,j,t)]\}|_P : i,j,t \in \{\text{S-T region}\}, \quad (1)$$

This feature is sensitive to the changes in overall amount of spatial activity within a given S-T region. For instance, localized blurring produce a reduction in the amount of spatial activity, whereas noise produces an increase of it.

Well-known [6] reference-free parameters describing the video sequence character are also the spatial

information (SI) and the temporal information (TI). SI is computed from the image gradient. It is an indicator of the amount of edges in the image:

$$\mathrm{SI} = \max_{\mathrm{time}_n} \{\mathrm{std}_{space}[\mathrm{Sobel}(F_n)]\}. \quad (2)$$

TI is computed from the pixel-wise difference between the successive frames:

$$\mathrm{M}_n(\mathrm{i,j}) = F_n^{(i,j)} - F_{n-1}^{(i,j)}. \quad (3)$$

It is an indicator of the amount of motion in the video:

$$\mathrm{TI} = \max_{\mathrm{time}_n} \{\mathrm{std}_{space}[M_n(i,j)]\}. \quad (4)$$

Finally, we investigated the codec compression settings frame rate (FR) and bit rate (BR), requiring no computational complexity for estimation as they are known at both sender and receiver.

IV. Data Analysis

For the reduction of the dimensionality of our data set while retaining as much information as possible, we used a well known multivariate statistical method, the Principal Component Analysis (PCA) [7]. The PCA was carried out to determine the relationship between MOS and the objective video parameters and to identify the objective parameters with the lowest mutual correlation, allowing us to propose a compact description of our data set. We performed PCA for all content classes separately. In our case first two components proved to be sufficient for an adequate modeling of the variance of the data (see Table II), because the total variability of the first two components was at least 81% for all content classes. Variability describes the percentual part of data sets variance that is covered by the variance of particular component. The horizontal axis represents principal component 1 (PC1) and vertical axis represents the principal component (PC2) at the figures 2, 3, 4, 5, 6. Each of the ten parameters is represented in the Figures 2, 3, 4, 5, 6 by a vector. The direction and length of the vector indicates how each parameter contributes to the two principal components in the graph.

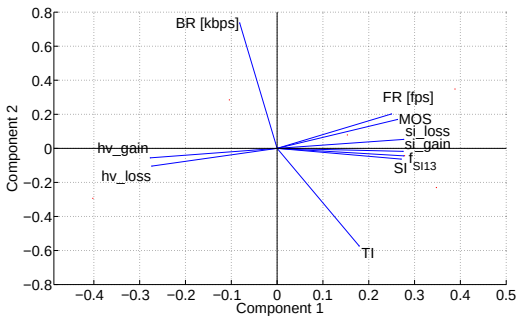

Fig. 2. Visualization of PCA results for ”akiyo” sequence.

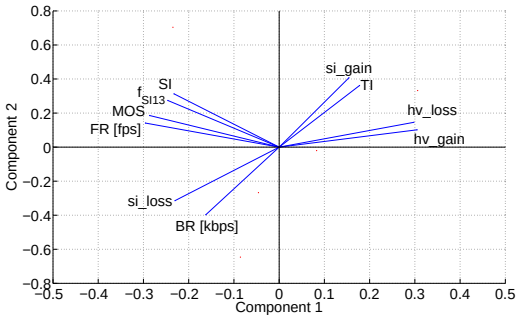

Fig. 3. Visualization of PCA results for ”foreman” sequence.

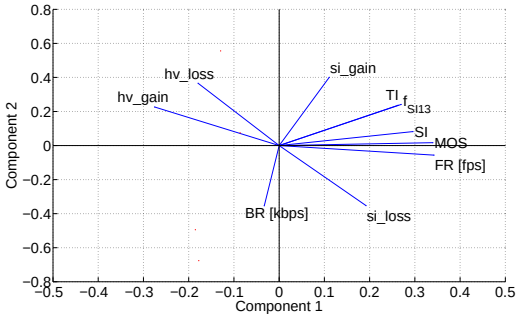

Fig. 4. Visualization of PCA results for ”soccer” sequence.

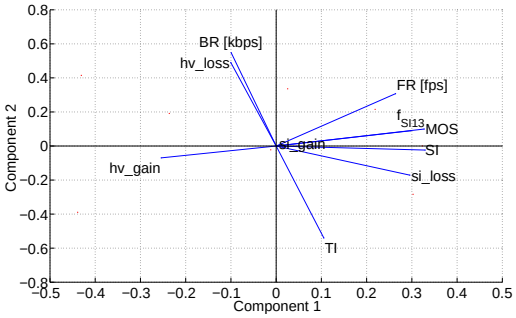

Fig. 5. Visualization of PCA results for ”panorama” sequence.

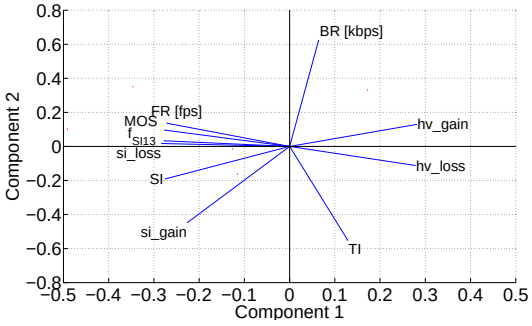

Fig. 6. Visualization of PCA results for ”traffic” sequence.

According to this analysis, each parameter in the analyzed data set can be assessed concerning its contribution to the overall distribution of the data set. This is

achieved by correlating the direction of the maximum spread of each variable with the direction of each principal component axis (eigenvector). High correlation between the first principal component (PC1) and investigated parameters indicate that the variable is associated with the direction of the maximum amount of variation in the data set. A strong correlation between the parameters and the second principal component (PC2) indicates that the variable is responsible for the next largest variation in the data, perpendicular to PC1.

Sequence	Variability of PC1 [%]	Variability of PC2 [%]
akiyo	84	14
foreman	67	28
soccer	55	37
panorama	81	10
traffic	61	20

TABLE II
THE TOTAL VARIABILITY OF THE FIRST TWO COMPONENTS FOR ALL CONTENT CLASSES.

Conversely, if a parameter (vector) does not correspond to any PC axis and its length is small compared to PC axis dimensions, this usually suggests that the variable has little or no control on the distribution of the data set. Therefore, PCA suggests, which parameters in our data set are important and which ones may be of little consequence. The parameters with the largest impact were taken into account metric design as is described in the following secton.

V. METRIC DESIGN

According to the PCA results we found the most suitable objective parameters for the metric design relevant for all content classes. Surprisingly, the objective parameters with higher complexity ($sigain$, $siloss$, $hvgain$, $hvloss$) did not correlate very well with PC1 and PC2 for all sequences. PCA results (see Figures 2, 3, 4, 5, 6) and high complexity of these objective parameters show us that these parameters are not appropriate for metric design in our scenario. If we take into account all content classes, complexity of objective parameters and PCA results, the most suitable parameters are FR, f_{SI13}, SI, TI and BR.

These experiences determine our further steps in metric design. We propose a metric having different coefficient values for different content classes because spatial and temporal sequence characteristics of our sequences are significantly different. We designed our metric using three objective parameters according to the correlation with PC and their complexity. We choose two parameters which correlate best with PC1, and one correlating with PC2 over all sequences because the variability of PC1 is approximately two times higher than the variability of PC2 (see Table II). Finally, we base our metric on the basic codec parameters FR and BR and sequence character parameter f_{SI13}. The uniform mathematical model for all content classes has been chosen due to its simplicity and rather good fit with the measured data. We improved the simple linear model with four mixed terms reflecting the most important mutual combinations of selected objective parameters.

Coeff.	Akiyo	Foreman	Fussball	Panorama	Traffic
K	$-78,4283$	$3,5970$	$-6,1850$	$-12,6834$	$-11,1982$
A	$-0,0302$	$0,0411$	$0,0241$	$0,0226$	$0,0322$
B	$0,1382$	$-0,0371$	$-0,0117$	$-0,0688$	$-0,02701$
C	$0,9252$	$-0,0288$	$0,1237$	$0,1429$	$0,14415$
r	$0,95$	$0,96$	$0,98$	$0,93$	$0,90$

TABLE III
COEFFICIENTS OF LINEAR METRIC MODEL AND CORRELATION OF AVERAGE MOS WITH SO OBTAINED ESTIMATION FOR ALL CONTENT CLASSES

Coeff.	Akiyo	Foreman	Fussball	Panorama	Traffic
K	$-39,0282$	$50,6525$	$98,3703$	$134,2721$	$-181,0251$
A	$2,1618$	$1,4769$	$-1,1826$	$-3,4122$	$2,2139$
B	$-3,2939$	$-4,9962$	$-13,7067$	$-33,2325$	$11,1780$
C	$0,4491$	$-0,5986$	$-1,7714$	$-1,2473$	$2,0426$
D	$-0,1338$	$-0,0936$	$0,1623$	$0,7717$	$-0,0859$
E	$-0,0234$	$-0,0158$	$0,0219$	$0,0325$	$-0,0244$
F	$0,0407$	$0,0592$	$0,2479$	$0,3134$	$-0,1256$
G	$0,0014$	$0,0010$	$-0,0029$	$-0,0073$	$0,0010$
r	$0,989$	$0,997$	$0,996$	$0,999$	$0,974$

TABLE IV
COEFFICIENTS OF IMPROVED METRIC MODEL AND CORRELATION OF AVERAGE MOS WITH SO OBTAINED ESTIMATION FOR ALL CONTENT CLASSES

$$\begin{aligned} MOS_v = \quad & K + A \cdot \mathrm{BR} + B \cdot \mathrm{FR} + C \cdot \mathrm{f_{SI13}} \\ & + D \cdot \mathrm{BR} \cdot \mathrm{FR} + E \cdot \mathrm{BR} \cdot \mathrm{f_{SI13}} \\ & + F \cdot \mathrm{FR} \cdot \mathrm{f_{SI13}} + G \cdot \mathrm{BR} \cdot \mathrm{FR} \cdot \mathrm{f_{SI13}}. \end{aligned} \tag{5}$$

To evaluate the quality of the fit of our proposed metrics for our data, we used a Pearson (linear) [9] correlation factor:

$$r = \frac{\mathbf{x}^T \mathbf{y}}{\sqrt{(\mathbf{x}^T \mathbf{x})(\mathbf{y}^T \mathbf{y})}}, \tag{6}$$

and the Spearman rank correlation factor [9]:

$$r' = 1 - \frac{6(\mathbf{x}-\mathbf{y})^T(\mathbf{x}-\mathbf{y})}{N(N^2-1)}. \tag{7}$$

In the case of metric model evaluation, vector $\mathbf{x}$ corresponds to **average** MOS values (averaged over three runs of all 38 subjective evaluations for particular test sequence) for all tested encoded sequences. Vector $\mathbf{y}$ corresponds to the prediction made by the proposed metric. As shown in Table III, even the linear model has already quite good fit. This confirms our choice of objective parameters. Furthermore, we can clearly see in Tables III and IV that the metric model coefficients are different for each content class. This suggests that our choice of the content classes is right and subjective video quality is content dependent as already resulted from [1], [8].

The performance of the subjective video quality estimation compared to the subjective quality data is summarized in Table V. The quality of prediction is characterized by a Pearson and Spearman rank order

Content type	Akiyo	Foreman	Fussball	Panorama	Traffic
r	0.8918	0.8423	0.8636	0.7512	0.7684
r'	0.8170	0.8298	0.8526	0.7345	0.7350

TABLE V

PREDICTION PERFORMANCE BY CORRELATION WITH MOS RESULTS IN DATASET

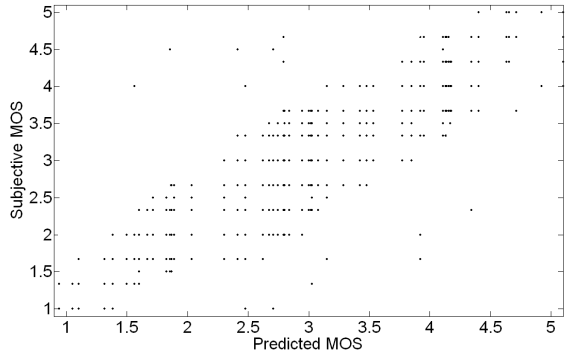

Fig. 7. Predicted vs. subjective MOS results.

correlations [9], where the vector $\mathbf{x}$ corresponds to the MOS data set values averaged over three runs of each test person and vector $\mathbf{y}$ to the metric prediction. Dimension of $\mathbf{x}$ and $\mathbf{y}$ refers to N.

The best correlation results were obtained for the first three sequences "akiyo", "foreman" and "soccer". The "akiyo" sequence can be compressed easily as it contains low amount of both spatial and temporal information. Therefore, for low BR we obtained very good MOS results. On the contrary, the test persons were extremely critical to the "soccer" sequence, because high compression ratio and low resolution leads in this case to very annoying effects, where the ball and the play field lines disappearing. The MOS for the "akiyo" and "soccer" sequences does not vary much and for these sequences we obtained noticeable fit. The "foreman" sequence was evaluated more critically than the sequence "akiyo". They are more sensitive to the FR. The entire sequence is of more dynamic nature than "akiyo", but does not contain as rapid movement as for example the "soccer". The "Panorama" and "traffic" sequences contain significantly lower amount of information. Therefore, the test persons do not use the whole MOS scale and it leads to lower fit of "panorama" and "traffic" sequences. For the "panorama" sequence, the quality of the still picture was most important. The best MOS evaluation obtained for lower FR [10]. The goodness of fit is at least 73% for Spearman and at least 75% for Pearson correlation for "panorama" sequence, althogh for the "akiyo", "foreman" and "soccer" we obtained fit above 81% for boths correlations.

The proposed metric design methodology we evaluated with mean value free correlation factor [7] defined as follows:

$$r'' = \frac{(\mathbf{x}-\overline{x})^T(\mathbf{y}-\overline{y})}{\sqrt{((\mathbf{x}-\overline{x})^T(\mathbf{x}-\overline{x}))((\mathbf{y}-\overline{y})^T(\mathbf{y}-\overline{y}))}}, \quad (8)$$

where the vector $\mathbf{x}$ corresponds to the **whole** MOS data set values averaged over three runs of each test person and vector $\mathbf{y}$ to the metrics prediction for all content types. Correlation coefficient lies between +1 and -1: a value of +1 implies a perfect association between $\mathbf{x}$ and $\mathbf{y}$, while a value of zero implies no association between $\mathbf{x}$ and $\mathbf{y}$ and value of -1 implies a perfect negative association between $\mathbf{x}$ and $\mathbf{y}$. The fitting performance of metric design methodology evaluated with (equation (8)) though all sequences and proposed metrics is 89%. This results clearly show that our proposed metrics and methodology are reliable for the video quality estimation.

VI. CONCLUSION

In this paper we investigated, proposed and compared perceptual quality metrics for typical content types for video streaming services in 3G networks. The investigated video content types have significantly different spatial and temporal sequence characteristics. The obtained MOS depends on the sequence character [1]. Therefore, we need content dependent metrics with different coefficients for each content type. Our proposed reference-free metric is a reliable prediction measurement tool of subjective video quality. The presented results clearly demonstrate that it is possible to estimate video quality without original sequence based on low-complexity objective parameters.

VII. ACKNOWLEDGMENT

The authors would like to thank mobilkom austria AG&Co KG for supporting their research and to Antitza Dantcheva for valuable help with the data collection. The views expressed in this paper are those of the authors and do not necessarily reflect the views within mobilkom austria AG&Co KG.

REFERENCES

[1] 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.

[2] 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.

[3] 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, Sept. 2003.

[4] 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.

[5] 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, Sept, 2004.

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

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

[8] M. Ries, R. Puglia, T. Tebaldi, O. Nemethova, M. Rupp, "Audiovisual Estimation for Mobile Streaming Services," Proc. of International Symposium on Wireless Communication Systems IEEE Ed., Sept, 2005.

[9] 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/.

[10] M. Ries, O. Nemethova, B. Badic, M. Rupp, "Assessment of H.264 Coded Panorama Sequences," Proc. of the First International Conference on Multimedia Services and Access Networks, Orlando, Florida, June 2005.