A Subjective Quality Estimation Tool for the Evaluation of Video Communication Systems

Jânio M. Monteiro, Mário S. Nunes

***Abstract*— This paper presents a quantification tool for the evaluation of scalable and non-scalable video communication systems. The proposed mechanism estimates the subjective quality of experience (QoE) of a human viewer according to the temporal resolution, the spatial resolution and the Root Mean Square of the Error (RMSE) between the original image and the encoded one. According to these three dimensions of quality it enables an encoder and server to search for the best combination of each of these scalability factors in order to deliver the best quality. The proposed quantification tool was obtained through subjective tests using a panel of evaluators and a new methodology which have shown good correlation factors between measurement data and estimating functions.**

***Index Terms*— Television over IP, H.264 Scalable Video Coding (SVC), Quality of Experience (QoE), subjective quality.**

I. Introduction

The popularity of Internet video streaming has grown tremendously as does the interest for mobile TV over IP with the growth of mobile terminals using 3G technologies and wireless LANs. In this field there are currently several mobile operators already delivering live encoded videos to clients with heterogeneous terminal capabilities.

In live video distribution, the traditional solution to adapt the encoded video to the receiver capability and bandwidth is simulcast. In simulcast several streams are encoded, targeted for different transmission rates and are delivered according to the bandwidth of the receiver. This solution, however, presents a limited number of quality profiles and it is necessary to frequently switch between encoded streams in order to search for the best quality.

An alternative to the discrete nature of simulcast is based in a layered encoding of video, supporting what is known as Fine Grained Scalability, which enables nearly continuous quality degradation according to the receiver bandwidth. In this field, in 2005 the MPEG and the Video Coding Experts Group (VCEG) of the ITU-T have joined efforts to define the Scalable Video Coding (SVC) as an Amendment of the H.264/MPEG-4 AVC standard [1]. SVC provides scalable video streams which are composed of a base layer and one or more enhancement layers. Enhancement layers may enhance the temporal resolution (i.e. frame rate), the spatial resolution (i.e. image size), or the signal-to-noise ratio resolution (SNR) of the content represented by the lower layers.

The reduction in temporal and spatial resolutions can hardly be measured in terms of traditional metrics and it is not yet completely known how each of these factors contributes to the perception of quality degradation in the specific scenarios of current video over IP or mobile terminals.

One of the differences between traditional television and mobile TV terminals rests in the type and size of terminals. Mobile terminals were developed to be used personally and therefore users expect reduced dimensions. Additionally the one-by-one usage of these terminals reduces the distance from the user and the display, and usually, the videos delivered present small sizes. It is therefore expectable that quality assessment evaluations should consider this type of scenarios.

The oldest method used to monitor video quality is subjective assessment. Subjective methods were standardized in ITU-R recommendation BT.500 [2] and have been used to evaluate video quality in television services for more than twenty years. However, subjective assessment tests previously designed to evaluate the quality in traditional television can hardly be applied to the current scenarios of television over IP (IPTV).

In [3] the authors propose a new subjective evaluation methodology called Subjective Assessment Methodology for Video Quality (SAMVIQ) and use it to evaluate the quality of several encoders. The advantage of this solution is that it was specially developed for Video over IP assessment.

In this paper we propose a subjective quality estimation tool for the continuous assessment of television over IP distribution systems. The proposed tool is meant to be used at the server side, in order to help deciding which arrangement of scalability factors best enhances the user perceived quality.

Although this methodology specifically focuses the SVC encoder it can also be extended to non-scalable video encoders.

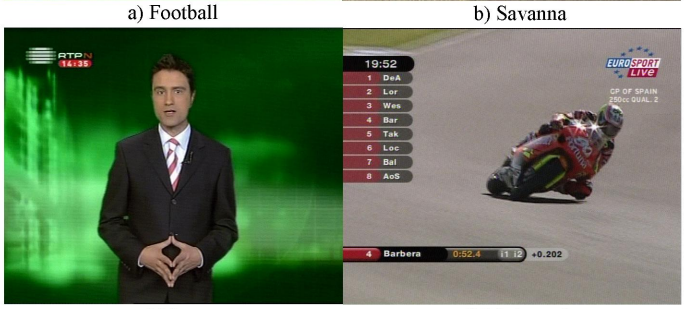
The following chapters are organised as follows. Chapter II describes the overall methodology used in the assessment tests. Chapters III and IV describe the procedures made and results taken from the assessment of the subjective quality, respectively as a function of the RMSE and as a function of spatial and temporal scalabilities. Chapter V defines a combined QoE expression and employs it in the analysis for the estimation of the quality of an SVC encoded video sequence. Finally Chapter VI concludes the paper.

J. M. Monteiro is with the University of Algarve, Faro, Portugal (email: jmmontei@ualg.pt) and INESC/IST, Lisboa, Portugal.

M. S. Nunes is with INESC/IST-UTL, Lisboa, Portugal (e-mail: mario.nunes@inov.pt).

1-4244-1521-7/07/$25.00 ©2007 IEEE

a) Football b) Savanna



c) News d) Motorcycle

Fig. 1. Sequences used in the metric tests.

II. Metric Tests Methodology and Video Sequences

A. Metric Tests Methodology

The value of the subjective quality (Q) is usually derived directly or indirectly through the Mean Opinion Score (MOS) of subjective assessment tests, like those proposed in the ITU-R BT.500 Recommendation. This recommendation considers and describes several metric tests, all them oriented to video quality assessment.

Besides the solutions proposed by the ITU-R BT.500 Recommendation, a new assessment test named SAMVIQ [3] was proposed in the European Broadcasting Union (EBU) Technical Review.

The SAMVIQ method allows the assessment of a set of videos (structured in several scenes) in a comparative way, grading each video in a linear scale between 0 and 100%. In each scene observers are firstly presented with a high anchor reference video (which must also be graded by observers) and all the other videos are presented in a random order. Similarly to other methods, this method defines the utilization of a hidden reference having the same quality as the reference video. In this methodology assessors can play each video as many times as they want in order to make a better judgement of the quality of each one.

Since users also grade the Reference video, the maximum value of the voting scale is known and can be afterwards used in the process of quality estimation. Particularly, that value can be used to compensate the scale boundary effect, using it as an anchor of maximum quality (i.e. the maximum scale limit used by observers).

This new method was selected in this study because it is more oriented to quality assessment of video transported over the Internet, being characterized as simpler and quicker than the previous ones.

In order to perform the tests, examiners (or subjects) were chosen among college students with ages ranging from 18 to 28 years old. All selected examiners were asked to answer an extensive questionnaire in order to validate the universe chosen for the video tests. The questionnaire also focused problems dealing with visual acuity in order to discriminate individuals with visual deficiencies.

B. Selected Video Sequences

Four video sequences with very different characteristics were used the measure the degradation of the perceived quality. This study was oriented to the evaluation of television equivalent videos and therefore all these sequences (represented in Fig. 1) were captured from TV broadcast sources. They were recorded using a frame rate of 25 fps, in the 4 times Common Intermediate format (4CIF, 704x576) and they were saved in an YUV format. Each video sequence was chosen in order to represent a different level of visual complexity in what refers to motion and image details.

Table I depicts the characteristics of these sequences and in which way each one relates to the sequences that are usually used in video quality metrics tests.

TABLE I
Video Sequences Characterization and Description.

Sequence	Most similar standard test sequence	Frames	Duration (seconds)
News	Akiyo	274	12.3
Football	Tempete	356	14.2
Motorcycle	Mobile	273	10.9
Savanna	Hall	281	11.3

Sequence News is characterized as having a fixed camera and presenting a full body newscaster over a synthetic background. Sequence Football is characterized for presenting a football match counter-strike with lots of detail and movement. Sequence Motorcycle is characterized for having rapid and complex details of an in road motorcycle grand prix. Sequence Savanna is a fixed camera in which movement is localized.

In all tests, the distance between each observer and their monitor ranged between approximately 0.4 and 0.6 m, and all monitors were set with the same resolution (1280x1024) and settings.

III. Quality as a Function of the RMSE

The Root Mean Square of the Error (RMSE) and the Peak Signal-to-Noise Ratio (PSNR), which can be derived from the RMSE, are currently two of the most common metrics used to assess compressed video quality.

There are however several studies, like for instance [4], which state that these metrics do not correlate well with the human perception of quality obtained through MOS derived from subjective tests. Particularly, it was verified that the human perception of quality is affected by many different artefacts introduced by compression algorithms, like for instance the blocking and ringing effects, which many times are not correctly quantified using the PSNR or RMSE metrics, since they contribute differently to subjective perception of

quality.

Because of this low correlation between objective and subjective metrics, several distinct solutions were proposed which attempt to estimate the subjective quality of video using other metrics that try to model the behaviour of the Human Visual System (HVS). Some of these solutions can be found on [5] and [6]. Nevertheless, none of these models have proven to be applicable to a wide range of circumstances. Instead, although they present a high degree of correlation in some conditions they also lead to low correlations under other circumstances.

In this context, it may be understandable that both, PSNR and RMSE, continue to be the most common metrics used to evaluate the quality of video compression algorithms and as a consequence of that, it would be important to find if, and in which cases, these metrics could be used to estimate subjective video quality. Based in these results, it would also be important to identify the mapping function that best estimates the HVS behaviour, evaluating it through standard procedures.

There are many advantages of converting PSNR or RMSE in subjective metrics. For instance it is commonly verified an increase in transmission rate of videos without a correspondence in perceptual quality. This usually happens because above a certain limit of encoding rate the effect in quality is almost negligible. However, at the encoder the PSNR metric continues to increase, giving the impression that there is a real increase in the perceived quality.

Nevertheless, in order to convert the RMSE or PSNR metric in a subjective metric, it is important to understand and distinguish psychological effects that mainly result from the chosen methodology, from psychological effects that estimates the Human Visual System response. Although the first ones usually cause distortions and variability in the results and for that reason need to be understood and compensated, the second ones should be preserved and correctly evaluated.

Additionally, since previous research have shown that different visual effects contribute differently to the perception of quality by the Human Visual System, the range of variation in visual effects should be restricted by only using a certain encoder algorithm. Based in this assumption, it should be possible to obtain a function that represents the way quality changes according to the RMSE variation (or PSNR).

Finally, the results obtained through assessment tests should be fitted using an interpolating function that consequently will represent the degradation function of a certain encoder (according to the main encoding parameters used). Nevertheless, according to the previously defined procedure it is important to analyse each encoder separately and not to extrapolate these results to other conditions.

According to these guidelines, this chapter specifies a method that can be used to estimate, with a high degree of correlation, the subjective quality of an encoded video and an associated parametric estimation function that translates the human visual system response to the RMSE variation. This estimation function can also contribute to the understanding of how artifacts degrade perception quality more quickly, when does it happen, and how they are associated with the increase in RMSE (or PSNR). Different curves obtained for different video sequences can also help in understanding how test sequences influence the perception of quality degradation. Additionally since the resulting estimation function describes quality degradation induced by the type of RMSE impairments of a specific encoder, applying the proposed method to several encoders permits the comparison, understanding and evaluation of how different encoding artefacts degrade quality, as the RMSE increases.

A. Specific Methodology

Each of the four videos was encoded in H.264 Baseline (BL) adjusted to the transmission bit-rates of 64, 128, 256 and 512 kbps using the CIF format. The RMSE between the original and the coded video was computed, applied only to the luminance component of the image. Table IV presents the RMSE results computed for each of the encoded videos.

TABLE II
RMSE RESULTS FOR EACH OF THE ENCODED VIDEOS.

Sequence	Bit Rate of the Encoded Videos (kbps)			
	64	128	256	512
Football	9.4524	6.3832	4.3756	3.1120
Savanna	6.3612	5.3770	4.3106	3.2177
News	4.4212	3.3888	2.7198	2.3771
Motorcycle	11.9823	7.6391	5.2124	3.7072

As previously explained, the subjective metrics tests were conducted using the SAMVIQ methodology and the obtained data were analyzed having the Recommendation ITU-R BT.500 [2] statistical treatment procedure as reference.

These metric testes were performed using a panel of twenty one examiners. The same set of video sequences was presented to the audience, allowing examiners to quantify the video quality in a linear scale ranging from 0 to 100%.

B. Results and Analysis

Table III, presents the obtained MOS of each sequence and the average MOS for the four sequences.

TABLE III
MOS RESULTS OF THE SEQUENCES CODED AT DIFFERENT BIT-RATES.

Sequence	Encoding bit-rate (kbps)				
	64	128	256	512	Reference
Football	0.354	0.612	0.630	0.710	0.760
Savanna	0.572	0.686	0.782	0.796	0.804
News	0.480	0.690	0.752	0.782	0.826
Motorcycle	0.266	0.552	0.694	0.808	0.832
Average	0.418	0.635	0.715	0.774	0.806

Fig. 2 presents the MOS obtained for each sequence and the 95% confidence interval for each metric result according to the procedure described in [2].

By analyzing Table III and Fig. 2 it becomes clear that the examiners tend to evaluate the intrinsic quality of each video

and not only the difference of quality, as measured by PSNR or RMSE metrics. This effect can be noticed by the difference of MOS values for the several reference video sequences.

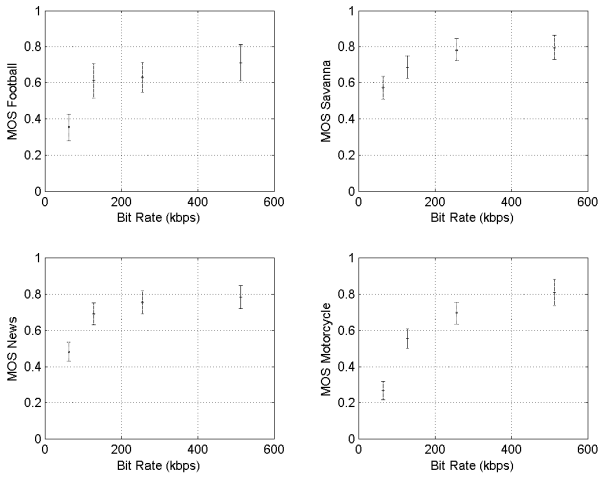


Fig. 2. Mean Opinion Scores and confidence intervals for each sequence encoded in H.264 BL.

In order to compensate this effect, all measurements were normalized using as reference the MOS obtained for the reference videos, as follows:

$$\overline{\overline{u}}_{jk} = \frac{\overline{u}_{jk}}{\overline{u}_{j_{REF}}} \quad (1)$$

were $\overline{\overline{u}}_{jk}$ represents the normalized value of the scorings for a certain sequence (for the test condition j, for a sequence k), $\overline{u}_{jk}$ is computed from the test videos (as defined in [2]) and $\overline{u}_{j_{REF}}$ represents the mean of scorings for the Reference videos of each sequence j.

Finally, the Normalized MOS values were interpolated using two functions. Although the recommended expression in [2] is:

$$Q = \frac{1}{1+e^{\theta(PSNR+\rho)}} \quad (2)$$

we found out through empirical analysis of data that the following function (3), besides being simpler, present higher values of correlation and shorter confidence intervals when compared with the previous one.

$$Q_{RMSE} = e^{-\alpha(RMSE)^2} \quad (3)$$

According to these two functions, the values of θ and ρ in function (2) and α in function (3) were computed.

TABLE IV
STATISTICAL PARAMETERS COMPARING THE EXPRESSION PROPOSED BY THE RECOMMENDATION ITU-R BT.500 AND THE ONE PROPOSED IN THIS PAPER.

Function	Coefficients	Standard Deviation	Correlation Coefficient
ITU-R BT.500 Recommendation	θ = -0.2619 ± 0.1006 ρ = -28.6823 ± 1.6800	0.0922	0.8764
Proposed in this paper	$\alpha = 8.05\text{x}10^{-3} \pm 2.00\text{x}10^{-3}$	0.0905	0.8833

Table IV presents these results along with the results of the standard deviation and the Pearson's correlation coefficient.

Fig. 3 presents the interpolation results using function (3), represented in a solid line. This curve translates the observer response to an increase in the error between the original image and the coded image, specifically taken for the H.264 encoder. In a dotted line is represented the 95% confidence interval of the function and in a dashed line is represented the 95% confidence interval for a new sample of Normalized MOS.

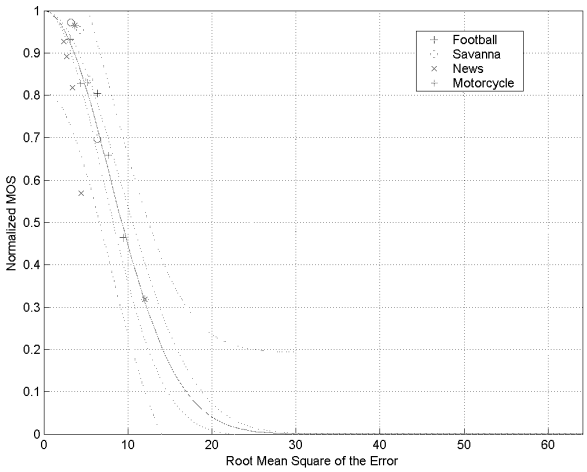


Fig. 3. Interpolation curve and values of the Normalized Mean Opinion Scores as a function of the Root Mean Square Error.

Although the obtained statistical results validate both, the methodology used and the selected interpolation functions, from the analysis of Fig. 3 it is noticeable that the encoding noise introduced in the News sequence degrades more quickly the perceived quality than the same amount of noise introduced in the other videos. This particular result is in line with previous studies like [4], showing that the type of noise being introduced by the encoding algorithm should be compared with the video characteristics in terms of texture.

IV. QUALITY AS A FUNCTION OF SPATIAL AND TEMPORAL RESOLUTIONS

Using the SAMVIQ quality assessment methodology, the subjective quality degradation introduced by reducing temporal and spatial scalabilities was quantified using the same four video sequences previously described, yet displayed uncompressed. These metric testes were performed using a panel of sixteen observers.

A. Specific Methodology

In this set of metric tests, observers were asked to evaluate the quality of video definitions for a 4CIF, CIF, QCIF and one quarter of QCIF (¼QCIF, 88x72) definitions and 25, 15, 7.5, 3.75 and 1.875 frames per second. All videos were displayed with the same size (as required by the SAMVIQ assessment procedure), with a 4CIF equivalent size (with an height of 15.2 cm). The 4CIF initial format was converted in the lower definitions (CIF, QCIF and ¼QCIF) using the default spatial

downsampling and resampling procedures available in the reference software of SVC [1].

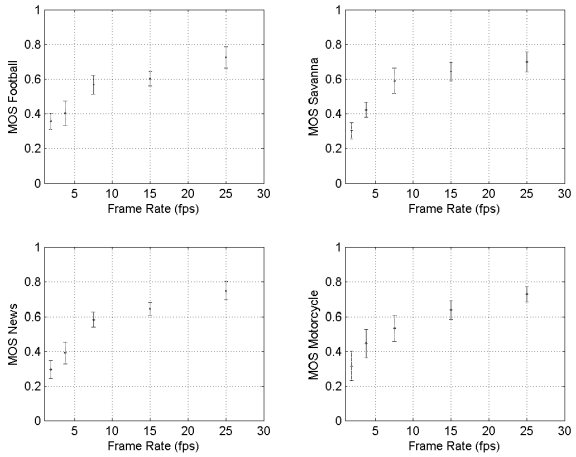


Fig. 4. Mean Opinion Scores and confidence intervals for each sequence with different frame rates.

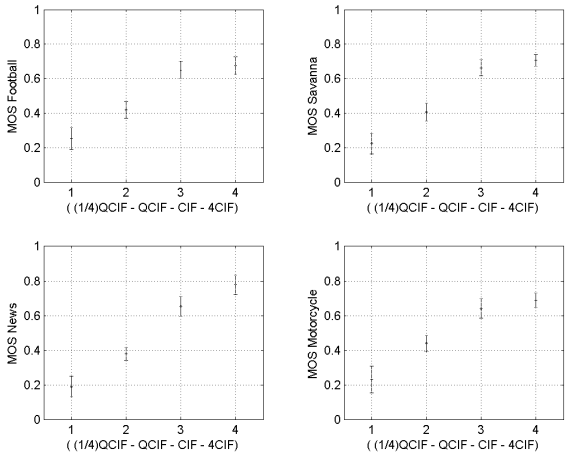


Fig. 5. Mean Opinion Scores and confidence intervals for each sequence with different definitions.

The image definition was evaluated in a linear scale between 1 (corresponding to the ¼QCIF format) and 4 (corresponding to the 4CIF). Intermediate values can be computed using the following expression:

$$Definition = \frac{1}{2}\log_2(PixelNumber) - 5.3147 \quad (4)$$

B. Results and Analysis

Figs. 4 and 5 represent the MOS obtained for each test sequence regarding the reduction in temporal and in spatial scalabilities, respectively.

In terms of temporal scalability the results have shown small differences in the MOS values between video sequences with different levels of movement.

Concerning Fig. 5, it can be verified that users tend to almost equally grade the quality of 4CIF and CIF sequences. However it is important to state that these results were oriented to the evaluation of television equivalent videos and therefore this difference could be higher if using as reference a higher definition video sequence.

The same normalization procedure described and implemented for the analysis of the impairments caused by the RMSE increase was implemented for these metrics (using expression (2)).

The Normalized MOS values were interpolated using the following functions (5) and (6), respectively for the reduction in temporal and the reduction in spatial scalabilities.

$$Q_{FrameRate} = \beta_1 + \beta_2 \log_{10}(FrameRate) \quad (5)$$

$$Q_{Definition} = \frac{1}{1+e^{-\delta_1(Definition-\delta_2)}} \quad (6)$$

The interpolation results are presented in Table V.

TABLE V
STATISTICAL PARAMETERS REGARDING THE CURVE INTERPOLATION ANALYSIS OF THE REDUCTION ON THE TEMPORAL AND SPATIAL SCALABILITIES.

Function	Coefficients	Standard Deviation	Correlation Coefficient
$Q_{FrameRate}$	$\beta_1 = 0.2827 \pm 0.0428$ $\beta_2 = 0.4634 \pm 0.0451$	0.0377	0.9813
$Q_{Definition}$	$\delta_1 = 1.1860 \pm 0.1648$ $\delta_2 = 1.8190 \pm 0.1115$	0.0424	0.9866

Fig. 6 presents both fitting functions (in solid line). In a dotted line is represented the 95% confidence interval of the function and in a dashed line is represented the 95% confidence interval for a new sample of Normalised MOS. As can be verified in Fig. 6, users are more tolerant to temporal scalability reduction than to spatial reduction.

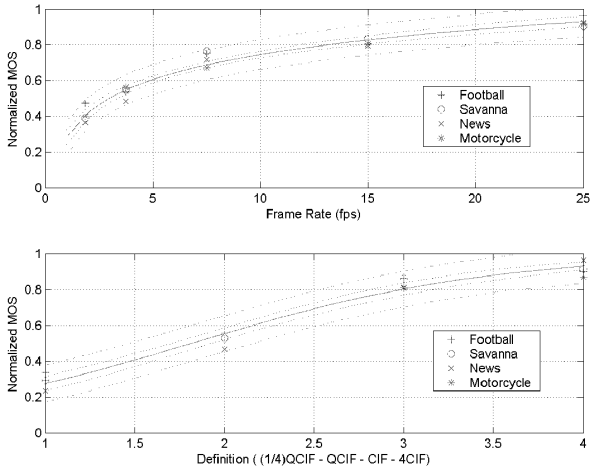


Fig. 6. Interpolation curves and values of the Normalized MOS with different temporal (upper part) and spatial (lower part) scalability measurements.

Both equations (5) and (6) were combined in a single function (7) expressing quality as a function of temporal and spatial scalabilities.

$$Q(FrameRate, Definition) = \gamma \frac{\beta_1 + \beta_2 \log_{10}(FrameRate)}{1+e^{-\delta_1(Definition-\delta_2)}} \quad (7)$$

In this expression, $\gamma = 1.0747$ was obtained through $\gamma = 1/max(Q_{FrameRate}, Q_{Definition})$ and serves to maintain the resulting scale within the maximum variation of both original

ranges.

Expression (7) is represented through a three dimensional plot in Fig. 7. Notice that these results are independent from the encoding algorithm since they were measured using raw videos, and therefore they can be applied to different encoders.

The next step is to join these results in a unique expression specifically applied to the H.264 encoder.

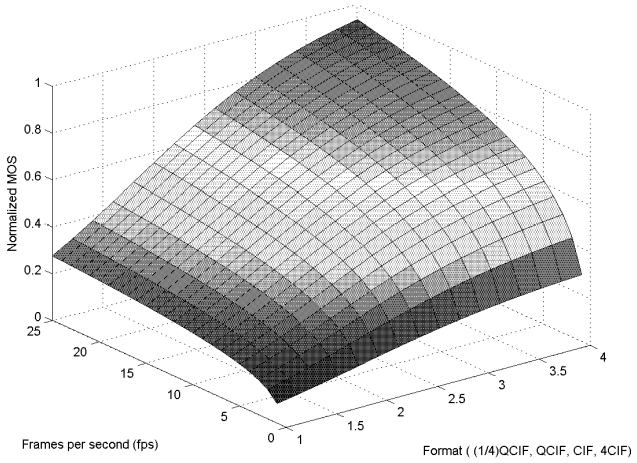


Fig. 7. Three dimensional representation of the Normalized Mean Opinion Score as a function of temporal and spatial scalabilities.

V. Combined Quality Expression and Analysis

Using the results obtained from Q_{RMSE} (which are specific to the H.264 encoder) with the results obtained from $Q_{FrameRate}$ and $Q_{Definition}$ it is possible to obtain an expression that estimates the HVS response to these three parameters. Since any of these quality factors should be capable of independently change quality in the full scale between 0% to its maximum level, we propose an expression that translates this behaviour which is obtained by multiplying both expressions (3) and (7), resulting in:

$$Q_T = \gamma e^{-\alpha(RMSE)^2}\left[\frac{\beta_1 + \beta_2 \log_{10}(FrameRate)}{1 + e^{-\delta_1(Definition - \delta_2)}}\right] \quad (8)$$

were Q_T refers to the total QoE of a viewer when facing these three scalability factors. This expression can also be rewritten as a function of the PSNR, since this metric can be directly obtained from the RMSE. The validity of expression (8) should be further evaluated using metric tests.

TABLE VI
Y-PSNR of the News sequence coded using the H.264 SVC encoder.

Format	Frame Rate (fps)				
	1.5625	3.125	6.25	12.5	25.0
QCIF	39.8 dB	39.2 dB	38.6 dB	38.3 dB	38.1 dB
CIF	40.4 dB	39.2 dB	38.3 dB	37.7 dB	37.5 dB
4CIF	39.6 dB	38.3 dB	37.6 dB	36.8 dB	36.4 dB

In order to demonstrate the applicability of expression (8), the News sequence was encoded using the reference SVC encoder [1]. Table VI show the Y-PSNR of the compressed video as a function of temporal and spatial scalabilities.

Fig. 8 represents the estimated QoE (for this encoding) as a function of the encoded bit rate. Fig. 8 shows that an increase in the encoding bit rate not always corresponds to an increase in user perceived quality and that a correct design of the quality enhancement layers should consider the associated subjective QoE factor here presented.

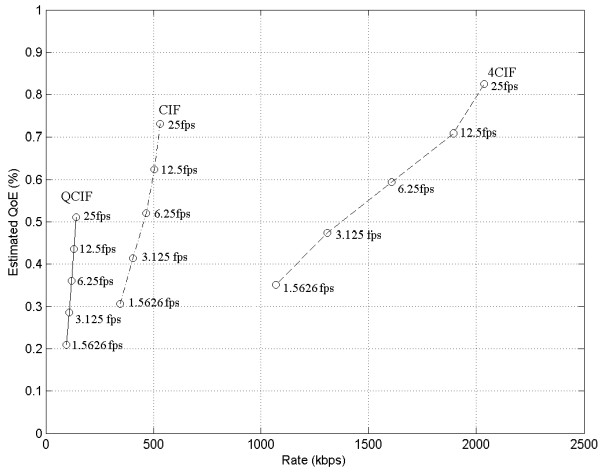


Fig. 8. Estimated QoE for the News sequence, coded using the SVC encoder, as a function of the encoding bit rate.

VI. Conclusion and Future Work

The results obtained in this paper enable the subjective quantification of the impairments caused by changing the temporal resolution, the spatial resolution and the RMSE in an encoded video. Using these results it is possible for an encoder (or video server) to select the best cost-benefit ratio, i.e. try to maximise the QoE according to the bandwidth available or the minimum bit rate for a specified level of quality.

These results also show that an increase in the Y-PSNR of an encoded video does not necessarily correspond to an increase in the subjective quality of a viewer. Therefore the estimated QoE must be considered as an important metric for the design and assessment of the encoder quality layers.

In terms of future work, further subjective tests should be performed in order to validate the proposed expression (8).

References

[1] T. Wiegand, G. Sullivan, J. Reichel, H. Schwarz, and M. Wien (eds.), "Scalable Video Coding – Joint Draft 9," Joint Video Team, Doc. JVT-V201, Marrakech, Morocco, January 13-19, 2007.

[2] ITU-R Recommendation BT.500-11, "Methodology for the subjective assessment of the quality of television pictures", June 2002.

[3] F. Kozamernik et al, "Subjective quality of internet video codecs — phase II evaluations using SAMVIQ", *European Broadcasting Union (EBU) Technical Review*, No. 301, January 2005.

[4] Girod, B., "What's wrong with Mean-Squared Error", Digital Images and Human Vision, A. B. Watson Ed., Chapter 15, pp. 207-220, MIT press, 1993.

[5] Guo, J. et al., "Gabor Difference Analysis of Digital Video Quality", *IEEE Transactions on Broadcasting*, Vol. 59, NO.3, Geneva, September 2004.

[6] Winkler, S., *Digital Video Quality*. John Wiley Ed., 2005, ch. 4.