A Study of Objective Quality Assessment Metrics for Video Codec Design and Evaluation

M. Martinez-Rach, O. Lopez, P. Piñol, M.P. Malumbres
Dept. of Physics and Computer Engineering
Miguel Hernández University
{mmrach,otoniel,pablop,mels}@umh.es

J. Oliver
Dept. of Computer Engineering
Technical University of Valencia
joliver@disca.upv.es

Abstract

When comparing the performance of different video coding approaches, improvements or new codec designs, one of the most important performance metrics is the Rate/Distortion (R/D), where distortion use to be measured in terms of PSNR (Peak Signal-to-Noise Ratio) values. However, it is well known that this metric not always capture the distortion perceived by the human being. So, a lot efforts were performed to define an objective video quality metric that is able to measure video quality distortion close to the one perceived for the destination user. In this work, we perform a study of different available objective quality metrics in order to evaluate their behaviour, taking as reference the classical PSNR metric. Our purpose is to find, if any, a video quality metric that is able to substitute PSNR for video quality assessment and determine a more accurate R/D performance metric when designing and evaluating video codec proposals.

1. Introduction

In the past years, the development of novel video coding technologies has spurred the interest in developing digital video communications. The definition of evaluation mechanisms to assess the video quality plays a major role in the overall design of video communication systems.

The most reliable way of assessing the quality of a video is subjective evaluation, because human beings are the ultimate receivers in most applications. The *Mean Opinion Score* (MOS), which is a subjective quality metric obtained from a number of human observers, has been regarded for many years as the most reliable form of quality measurement. However, the MOS method is too cumbersome, slow and expensive for most applications. The *objective* quality metrics are valuable because they provide video designers and standards organizations with means for making meaningful quality evaluations without convening viewer panels. So, the objective will be to find an objective quality metric that exhibits a good behaviour for a large set of video distortions and, what it is most important, get measures as much as close to the ones perceived by human observers. Also it would be desirable that the time required for giving a quality measure will be short enough for their practical use.

In the literature, there is a consensus in a primer classification of objective quality metrics [1][2] attending to the availability of original non-distorted info (video reference) to measure the quality degradation of an available distorted version:

Full Reference (FR) metrics perform the distortion measure with a full access to the original image/video version which it is taken as a perfect reference.

No Reference (NR) metrics have no access to reference image/video. So, they have to perform the distortion estimation only from the distorted version. In general they have lower complexity but are less accurate than *FR* metrics and are designed for a limited set of distortions and video formats.

Reduced Reference (RR) metrics work with some information about the original video (similar to a perceptual hash algorithm). An *RR* metric defines what information have to be extracted form original video, so it can be compared with the one extracted in the distorted version.

The most widely used *FR* objective video quality metrics by the scientific community are Mean Square Error (MSE) and PSNR. They are simple to calculate, and mathematically easy to deal for optimization purposes providing an easy way to evaluate the video quality [3]. However, it is well known that not always capture the distortion perceived by the Human Visual System (HVS)

In the last years, new objective image and video quality metrics have been proposed in the literature, mostly for FR/RR Quality Assessment (QA). They emulate human perception of video quality since they

Proceedings of the Eighth IEEE International Symposium on Multimedia (ISM'06)
0-7695-2746-9/06 $20.00 © 2006 IEEE

produce results which are very similar to those obtained from subjective methods. Most of these proposals were tested in the different phases carried out by the Video Quality Experts Group (VQEG) which was formed to develop, validate and standardize new objective measurement methods for video quality. Although the Phase I test [4], for FR television video QA only achieved limited success, VQEG continues its work on Phase II [5] test for FR QA for television, and RR and NR QA for television and multimedia.

In this work we are going to evaluate different available objective quality metrics to find candidates to replace the classical PSNR metric when different video coding proposals are evaluated by means of the R/D performance index. So, we will use a set of video encoders and video sequences in order to create Hypothetical Reference Circuits (HRC) and compare the QA results of the different objective quality metrics under study. Also, we will consider their complexity in order to determine their application area.

The organization of the paper is as follows: In the next section we will describe the main frameworks defined around objective QA metrics. In section 3, we describe the metrics and methods used for comparing objective quality metrics. In Section 4 we show the behaviour of several available quality metrics, including PSNR as reference. Finally, in section 5 some conclusions are given.

2. Objective quality metric frameworks

We have found in the literature different frameworks that group several metrics depending on the way they are designed. In this section we will briefly describe the main ideas behind the different frameworks and their main objective quality metrics.

2.1 Error Sensitivity

The Error Sensitivity framework (ESF) group all the metrics that were designed taking into account different models based on the current knowledge of the Human Visual System (HVS). Generally, the emulation of HVS is a bottom-up approach that follows the first retina processing steps to continue with different models about the visual cortex behaviour. Also, some metrics deal with cognitive issues about the human visual processing.

Usually the HVS models first decompose the input signal into spatio-temporal subbands in both the reference and distorted signal. Then, an error normalization and weighting process is carried out in order to give the estimated degradation measure.

Most metrics based on ESF are FR by definition. The main difference between them is related with the way they perform the subband decomposition inspired in the complex HVS models [6-8], low cost decompositions in DCT [9] or Wavelet [10] domains, and with other HVS related issues like in [11] where foveal vision is also taken into account.

2.2 Structural Distortion/Similarity

The *Structural Distortion/Similarity Framework* (SDF) is focused on a top-down approach, analyzing HVS to emulate it at a higher abstraction level. So, authors supporting this framework argument that the main function of the human eyes is to extract structural information from the viewing field, being the HVS highly adapted for this purpose. Therefore, a measurement of structural distortion should be a good approximation of perceived image distortion.

So it is assumed that the HVS does not perceives the quality of a visual scene as a function based on intensity and contrast variability. Instead of that, this framework will look for structural information that will be perceived at cognitive levels of HVS. Changes in contrast and luminance are not considered as modifications in the image structure. So, these metrics are able to distinguish two types of distortions: The ones that change the image structure and those distortions that do not change it. In [12] an image quality index is defined which is refined and improved in [13]. Also, in [14] the authors propose a generalization of their work where every distortion may be decomposed in a lineal combination of different distortion components. In [15] the model is extended to the complex wavelet domain in order to design a robust metric to scaling, rotation and translation effects.

In [16] a video quality metric is proposed following a frame by frame basis. It takes quality measures for different blocks of each frame taking into account their spatial variability and also weighting the movement and other effects (like blocking) by means of an specifically adapted NR metric [17].

2.3 Statistics of natural images

The third franework is related with the statistical behaviour of natural images and we will refer it as *Statistics of Natural Images Framework (SNI)*. In this framework a natural image/video is defined as those captured with high quality devices working in the visual spectrum (natural scenes). So, text images, computer generated graphics, animations, draws, random noise or image and videos captured with non visual stimuli devices like Radar, Sonar, X-Ray, etc. are out of the scope of this framework.

Authors supporting this framework argument that the HVS has evolved with the statistical patterns (spatial and temporal) found in the signals captured form the visual field. Also, they state that these statistical patterns of natural scenes have modulated the biological system, adapting the different processing layers to these statistics.

So, the metrics defined under this framework will extract the relevant information from visual input signal in form of statistical information. In [18] a statistical model of a wavelet coefficient decomposition is proposed, and in [19] the authors propose an NR metric derived from previous work.

In this framework, the distortions are defined as the ones whose statistic patterns are far away from the ones found in “perfect natural images”. In fact, some metrics defined under this framework take the objective quality assessment as an information lose problem, using approaches close to the information theory [20,21].

2.4 Other objective quality metrics

Finally, there are other metrics that we have not classified under the frameworks mentioned above and we will classify them in a *Specific Metric Framework* (SMF). Among them we can find metrics that valuate spatial information loses, edge shifting, and luminance and colour variability [22]. Also, we can find metrics based on watermarking techniques that analyze the quality degradation of the embedded image [23]. There are metrics that are designed for measure specific distortions types or the ones produced by specific encoders [24,25].

3. Metrics and Methods

We will briefly introduce the metrics we have found available for our study and the method we carried out to obtain a quality value in DMOS space (Differences Mean Opinion Score). QA Metrics under study are:

- Mean Structural SIMilarity index (MSSIM[1]) [26] a FR-Image metric from the SDF. In the reference paper, the metric was tested against JPEG and JPEG2000 distortion types, but we include the new distortion types available in the new release of Live database[2] because the aim of the structural approach is to be a general one.
- Visual Information Fidelity (VIF[3]) measure [27] located in the SNI framework, a FR-Image metric that acts as an image information measure that quantifies the information that is present in the reference image, and also quantifying how much of this reference information can be extracted from the distorted image.
- No-Reference JPEG Quality Score (NRJPEGQS[4]) [24] a NR-Image metric designed specifically for JPEG compressed images. Extracts features that can be used to reflect the relative magnitudes of blocking and blurring combined to constitute a quality prediction model.
- No-Reference JPEG2000 Quality Assessment (NRJPEG2000[5]) [16] a NR-metric that use Natural Scene Statistics models to quantify the departure of a distorted image from "expected" natural behaviour.
- Reduced-Reference Image Quality Assessment (RRIQA[6]) [20] the only RR-metric under study which is based on a Natural Image Statistic model in the wavelet transform domain and use the Kullback-Leibler distance between the marginal probability distributions of wavelet coefficients of the reference and distorted images as a measure of image distortion.
- Video Quality Metric (General Model) (VQM[7]) [22] is a video FR-metric adopted as standard by the American National Standards Institute (ANSI) in 2003. The International Telecommunication Union (ITU) has also included the NTIA General Model as a normative method in two Draft Recommendations.
- The traditional PSNR in the predicted DMOS Space, that we call DMOSp-PSNR.

Each QA Metric scores the quality of the image/video using an specific scale. In order to compare the behaviour of various metrics for a set of images/sequences, the objective quality index obtained for each metric has to be converted into a common scale. We will use a non-linear parametric mapping function to convert the objective quality index of each metric to the common Predicted-DMOS space (DMOSp). The mapping of the quality index of metrics to the subjective scores depends on the methodology, validation and application scope of the subjective tests. Therefore it is not included in the QA algorithm and it is usually done by the final application that use the metric. In the VQEG Phase-I and Phase-II testing and validation [4,5], and in other extensive metrics comparison tests [28], a non-linear mapping between the objective and the subjective scores was allowed, and the performance validation metrics are computed after a non-linear curve fitting [29].

[1] http://www.cns.nyu.edu/%7Ezwang/files/research/ssim/index.html
[2] http://live.ece.utexas.edu/research/quality/subjective.htm
[3] http://live.ece.utexas.edu/research/Quality/VIF.htm
[4] http://www.cns.nyu.edu/%7Ezwang/files/research/nr_jpeg_quality
[5] http://live.ece.utexas.edu/research/Quality/nrqa.htm#nrqajpeg2000
[6] http://www.cns.nyu.edu/%7Ezwang/files/research/rriqa/index.html
[7] http://www.its.bldrdoc.gov/n3/video/vqmsoftware.htm

$$Quality(x) = \beta_1 \text{logistic}(\beta_2, (x\text{-}\beta_3)) + \beta_4 + \beta_5 \qquad (1)$$

$$\text{logistic}(\tau, x) = \frac{1}{2} - \frac{1}{1+\exp(\tau x)} \qquad (2)$$

The common value space used for comparing the performance of the metrics in this study is DMOS (Differences Mean Opinion Score). Another useful scale could be JND which has a better inherent meaning than DMOS and is not subject to criterion and context effects [33]. We choose for this work the DMOS scale because of the availability of DMOS values in the used image/sequence databases. Raw scores obtained in subjective tests are converted into difference scores and processed further [21] to get a linear scale in the 0-100 range, where 0 represents the best quality value (no difference between reference and distorted image).

Once the subjective scores of image/video are available is time to run each metric under test. For FR-metrics both reference and distorted images/videos are the input, for NR-metrics only distorted image/video and for RR-metrics the reference image/video is the input of the features extraction step and the extracted features and the distorted image/video are the input for the final metric evaluation step.

Each metric has to be trained with images/videos having the impairments for which was designed to handle with, and then it will work with another image/video set that we call *'test set'*. So in our study SSIM, VIF, RRIQA and DMOSp-PSNR are trained with the whole Live2 database, NRJPEGQS is trained only with the JPEG distorted images of Live2 database, NRJPEG2000 is trained only with the JP2K distorted images of Live2 database and VQM-GM is trained with a subset of 8 video sequences and its 9 corresponding HRCs of VQEG Phase I database in the range of 1 to 4Mb/s bit-rate. All the metrics have been trained only with the information of the luminance component.

Sequence	Frame	F.Num.	F.Rate
Foreman	QCIF : 176 x 144	300	30 fps
Container			
Foreman	CIF : 352 x 288		
Container			
Mobile	CCIR*: 640 x 512	40	

Table 1. Sequences included in the *'test set'*

Having the objective quality indexes for all the HRCs and their corresponding subjective quality indexes, the next step is to get the parameters of Eq. 1 through a non linear mapping between objective and subjective scores.

The *'test set'* used comprise different standard video sequence used in video coding evaluation as shown in Table 1, using only the luminance component.

Finally for each metric and HRC in the *'test set'*, we will use Eq. 1 to obtain the correspondent DMOSp values (predicted DMOS). Image metrics were applied to each frame of the sequences and the mean objective quality for all the frames was translated to DMOSp.

We have measured the computation time needed for each metric (except for VQM-GM) to calculate its objective quality value for each frame in sequences at different frame sizes, and the mean value of the whole sequence is taken as time performance metric for the reference software of each metric.

4. Analyzing Results

In this section, we will proceed with the evaluation study, remarking that our purpose is to evaluate video codecs and to find out if there is a metric that could substitute the traditional PSNR to obtain more accurate R/D performance indexes in the process of design and evaluation new video encoding proposals.

	β1	β2	β3	β4	β5
MSSIM	-39.5158	14.9435	0.8684	-10.8913	46.4555
VIF	-3607.3040	-0.5197	-1.6034	-476.0144	-693.3585
NRJPEGQS	37.6531	-0.9171	6.6930	-0.2354	40.7253
NRJPEG2000	37.3923	0.8190	0.6011	-0.8882	74.5031
RRIQA	-18.9995	1.5041	3.0368	6.4301	5.0446
PSNR-PMOSp	23.2897	-0.4282	28.7096	-0.6657	61.5160
VQM-GM	-163.6308	6.3746	-7.6192	114.4685	76.6525

Table 2. Metrics equation 1 parameters.

We have used an Intel® Pentium® 4 CPU Dual Core 3.00, 3.00 GHz with 1 Gbyte RAM. The programming environment used is Matlab 6.5 Rel.13 (The MathWorks, Inc.). The Matlab source code of evaluated metrics is public available on the internet or supplied by the authors. The codecs under test are H.2647AVC [30], a DCT based codec running in intra and inter mode and two wavelet based image codecs, Motion-JPEG2000 [31] and Motion-LTW [32].

	CC	RMSE	SROCC
MSSIM	0,8625	7,9682	0,8510
VIF	0,9529	0,0516	0,9528
NRJPEGQS	0,9360	3,0837	0,9020
NRJPEG2000	0,9099	7,0560	0,9021
RRIQA	0,9175	4,9486	0,9194
PSNR-DMOSp	0,8257	9,0969	0,8197
VQM-GM	0,8957	7,6746	0,9021

Table 3. Goodness of fit DMOSp – DMOS

The fitting between objective metric values and subjective DMOS scores was done using the Matlab curve fitting toolbox looking for the best fit in each case. Performance validation parameters between DMOS and predicted DMOS values are Pearson

Correlation Coefficient (CC), Root Mean Squared Error (RMSE) and Spearman Rank Order Correlation Coefficient (SROCC). The betas for our fittings are shown in Table 2 and Table 3 shows the performance validation parameters.

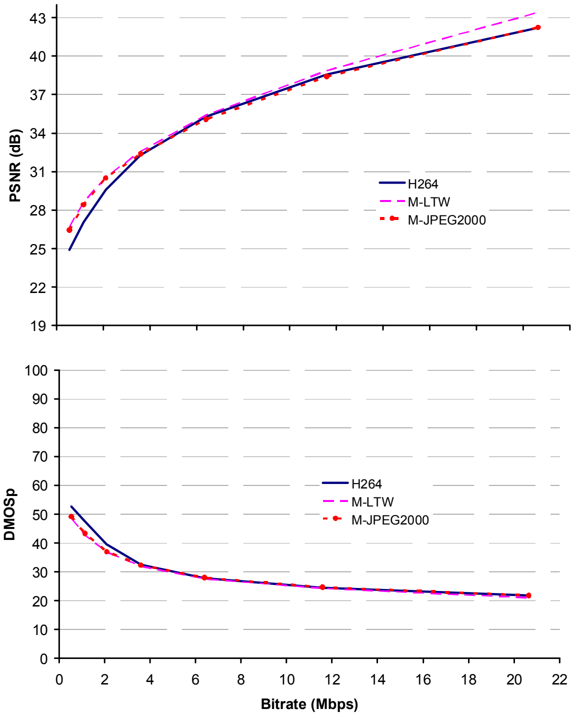

Figure 1. PSNR vs. DMOSp-PSNR for the evaluated codecs and for Mobile

A R/D plot of the different video codecs under test using the traditional PSNR as distortion measure is shown in upper panel of Figure 2. It is usual to evaluate performance of video codecs in a dynamic range varying from 20-22 dB to 40-44 dB but over 38-40 dB its difficult determine which one is better. This saturation effect at high qualities is not captured by the traditional PSNR, see upper panel of Figure 2.

We convert traditional PSNR to metric DMOSp-PSNR applying the corresponding betas in Eq. 1. We can see in lower panel of the subjective saturation effect above a specific quality for the DMOSp-PSNR metric. At bit-rates in the range from 11.5 Mbps to 20.5 Mbps the DMOSp values practically do not change. For all the evaluated codecs this behaviour is the same, and for all evaluated frame sizes increasing smoothly the slope of the saturation line as the frame size increase. This saturation effect agree with the fact that there is almost no noticeable subjective difference when watching the sequences at the two highest bit-rates. At the highest frame size evaluated, the slope for the DMOSp-PSNR metric gives differences from 2.66 to 3.28 DMOSp depending on the codec and this DMOSp variation range could be assumed as imperceptible.

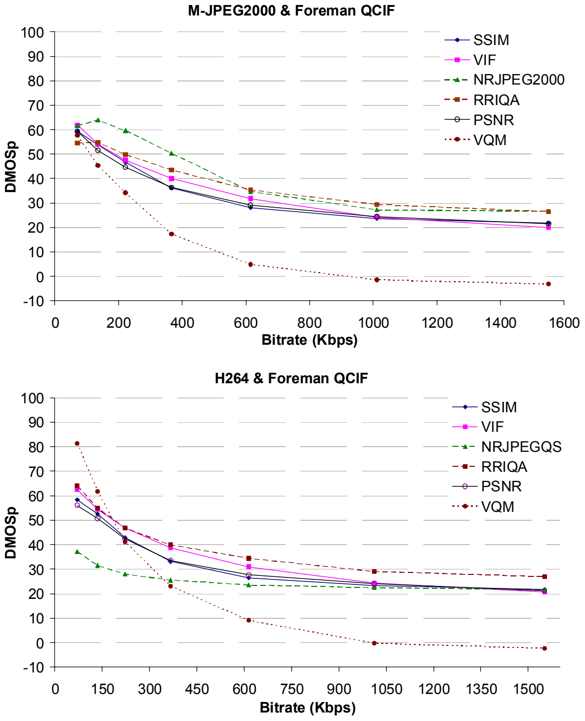

Figure 2. Codecs vs Sequences R/D plots

Figure 1 shows that at lowest bit-rate, the ranking quality order for the different codecs remains the same than for traditional PSNR and for DMOSp-PSNR. This behaviour remains for all sequences and for lower bit-rates than the bit-rate where the saturation effect begins with almost the same distances for the quality axis. This allow us to take the DMOSp-PSNR metric as the 'subjective' counterpart of the traditional PSNR when comparing these codecs at different bit-rates.

As PSNR is not a good perceptual metric for image or video quality assessment, now we look if the remaining metrics under study have the same behaviour, for low and high bit-rates, but with a best perceptual scoring.

Figure 2 shows some of the resulting R/D plottings used for comparing all metrics. The saturation effect is captured by all metrics at high bit-rates regardless the codec-sequence evaluated. There is almost no subjective noticeable differences at the two highest bit-rates. It could be thought that differences below 5 DMOSp values are not noticeable.

All metrics gives, as expected, a decreasing score of DMOSp when the bit-rate decrease. Looking at lower panel of Figure 2 and at the lowest bit-rate the DMOSp rating differences between metrics arrives surprisingly up to 44.21 DMOSp units. As shown in lower panel there are three different behaviours, VQM which was trained with VQEG sequences, NRJPEGQS trained only with JPEG distorted images and the rest of the metrics with all Live2 database distorted images.

Without having any subjective score available it is difficult to say which metric scores better increments in DMOSp between two consecutive bit-rates (according with subjective perception). This variations goes from 0.82 to 4.91 DMOSp for the processed sequences and codecs. As we can see, the DMOSp range that could be taken as imperceptible, depends on many factors (codec, frame size and metric), growing the mean differences as the frame size does. Besides it has been subjectively observed that the same variation in DMOSp is perceived, along the dynamic range of bit-rates, with different intensities.

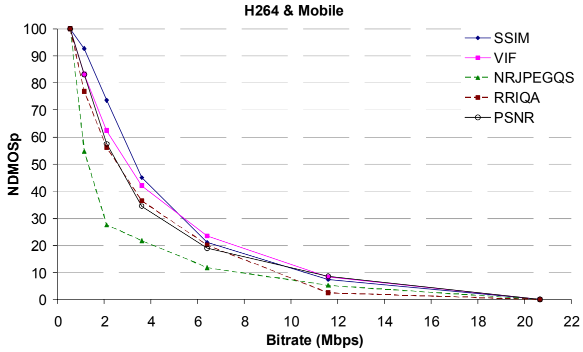

Figure 3. Normalized DMOSp values for metrics in a R/D plot

Normalizing DMOSp values by the dynamic range of each metric in a plot, and translating it linearly again to a 0-100 scale we get R/D plots in a Normalized DMOSp space (NDMOSp), Figure 3. Differences in this NDMOSp space have the same perceptual meaning regardless of the metric.

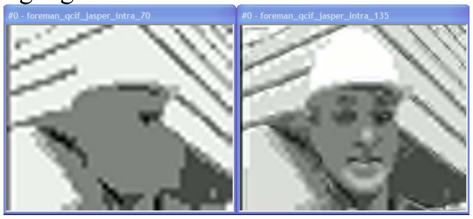
#0 - foreman_qcif_jasper_intra_70
#0 - foreman_qcif_jasper_intra_135

Figure 4. First frame of the foreman qcif at two consecutive bit-rates

Between the two highest bit-rates the biggest difference in NDMOSp is 8.62 that we appreciate subjectively as imperceptible. NRJPEGQS gives a NDMOSp difference of 5.83 (between 2.1 and 3.5 Mbps) and MSSIM gives a difference of 7.29 (between 0.54 and 1.14 Mbps). Therefore these metrics are reporting less differences that the one we know as imperceptible (at these bit-rates) but subjectively distortions are perceived.

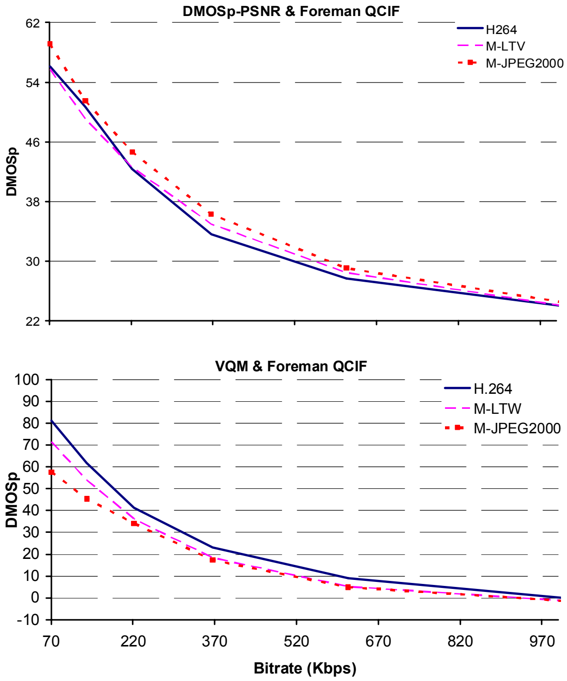

Figure 5. PSNR-DMOSp vs VQM. Ranking Codecs against Bitrates

Another alterations in the 'normal' behaviour of metrics when evaluating R/D performance plots are noticed. In the upper panel of Figure 2 and at the two lowest bit-rates the quality score of RRIQA and NRJPEG2000 decrease as the bit-rate increase, instead of increasing.

Figure 4 shows the first frame of these bit-rates. It is common to classify the right image (135 Kbps) better than left one (70 Kbps), not like RRIQA and NRJPEG2000. This only happens with M-JPEG2000, for RRIQA with Foreman QCIF, and for NRJPEG2000 with all QCIF and CIF sequences.

VQM at low bitrates changes the subjective ranking of quality between codecs before saturation. This subjective ranking (in descending quality for CIF is M-LTW, M-JPEG2000, H264 and for QCIF is M-LTW, H264, M-JPEG2000) agrees with the one given by DMOSp-PSNR at bit-rates before saturation, as shown in Figure 5 where the ranking for VQM changes.

Concerning the metrics trained with the same set, our performance validation data says that the metric who best fit to DMOS is VIF. In Figure 2 we see that the remaining metrics follows very close the scores of VIF along the bit-rate range regardless of the codec.

Up to now we have been analyzing results when codecs runs in intra mode. Now we will focus on the results obtained for H264 codec running in inter mode with the default settings.

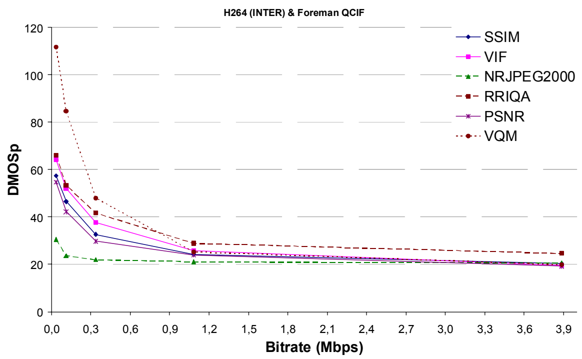

Figure 6. Behaviour of metrics when codec runs in inter mode

The behaviour for every metric as the bit-rate increase is the same as in intra mode, keeping the relative ordering of metrics. VQM sets the saturation quality approximately at the same DMOSp value as the rest of the metrics as shown in Figure 6. At lowest bit-rates, objective quality value of VQM falls out of the training range giving a DMOSp value over the maximum. NRJPEG2000 reacts as in intra mode, slowly as bit-rate decreases.

	QCIF		CIF		CCIR*	
	Frame	Seq	CIF	Seq	CCIR*	Seq
MSSIM	0,028	8,4	0,147	44,1	0,764	30,5
VIF	0,347	104,1	1,522	456,5	6,198	247,9
NRJPEGQS	0,010	3,0	0,049	14,6	0,201	8,1
NRJPEG2000	0,163	48,9	0,486	145,9	1,595	63,8
RRIQA (f.e.)	4,779	1433,7	6,950	2084,9	10,111	404,5
RRIQA (eval.)	0,201	60,2	0,635	190,6	2,535	101,4
PSNR	0,001	0,3	0,006	1,7	0,020	0,8

Table 4. Frame mean evaluation time and sequence evaluation time (seconds)

Finally, Table 4 shows for different frame sizes the frame mean evaluation time and the whole sequence evaluation time. Times for the two steps of RRIQA, features extraction (f.e.) and quality evaluation (eval.) have been separately measured. Times for VQM have been measured manually. For a CIF sequence VQM takes from 27 to 28 seconds (calibration and colour conversion time not included) which is faster than the other FR metrics except NRJPEGQS and DMOSp-PSNR. DMOSp-PSNR is far away the less computational expensive metric at all frame sizes. On the other hand, RRIQA and VIF are the slowest metric (they run a linear multi-scale, multi-orientation image decomposition) but they are the most accurate of the no distortion specific metrics.

5. Conclusions

In this work we have analyzed the comparison process of three video codecs, the DCT based H264 working in intra and inter mode and two motion implementation of wavelet based codecs, Motion-JPEG2000 and Motion-LTW (only intra mode) using public available Objective Quality Assessment Metrics. The main aim was finding a Quality Assessment Metric that can be used instead PSNR to achieve better adjustments to human perception of quality when valuating compressed video sequences at different bit- rates.

Metrics have to be compared in a common quality space. We used predicted DMOS (DMOSp) space. When comparing in the DMOSp scale is preferable do it with metrics trained with the same set. A R/D comparison of different kind of metrics (trained with different sets) must be done carefully, looking not only to the absolute quality scores but also to the degree that different metrics score the subjective differences between consecutive bit-rate variations. When metrics are trained with the same training set (differences in DMOSp values have the same perceptual meaning for all metrics), it can be trust the quality given by the metric which has better fit to DMOS in its calibration process.

Our results show that NRJPEG2000 gave wrong quality scores between the two highest compressed sequences with M-JPEG2000 codec in all sequences. RRIQA also failed with this codec but only for small frame sizes. NRJPEGQS metric is slow in perceiving the decreasing of quality and between some consecutive bit-rates does not perceive differences of quality as others metrics and subjective tests do. VQM ranks in bad order the codec performance for QCIF and CIF frame sizes. All metrics capture the saturation effect in perceived quality at high bit-rates.

In general each metric can be use depending on the application, the frame size, the bit-rate range used.

If there is no availability of the reference sequence RRIQA is our choice because has practically the same behaviour than FR metrics.

If the reference sequence is available the choice depends on the weight given to the trade-off between computational power and accuracy. If time is the most important parameter we will choose DMOSp-PSNR followed by VQM, and if accuracy is most important, then the choose will be VIF.

6. References

[1] S. Winkler, “Issues in vision modeling for perceptual video quality assessment”, Signal Processing, Elsevier, n.78, pp. 231-252, 1999.

[2] Z. Wang, H.R. Sheikh, A.C. Bovik “Objective Video Quality Assessment” Chap. 41 in The Handbook of Video Databases: Design and Applications, B.Furht and O.

Marqure, ed., CRC Press, pp. 1041-1078, Sep. 2003

[3] B. Girod, “What's wrong with mean-squared error”. Digital Im-ages and Human Vision, A. B. Watson, ed., pp. 207-220, MIT Press, 1993.

[4] VQEG, “Final report from the video quality experts group on the validation of objective models of video quality assessment. Phase I”. Mar. 2000. http://www.vqeg.org/.

[5] VQEG, “Final report from the video quality experts group on the validation of objective models of video quality assessment. Phase II”. August 2003. http://www.vqeg.org

[6] P.C. Teo and D.J. Heeger, “Perceptual image distortion”, Human Vision Visual Processing and Digital Display V_ IST&SPIE Symposium on Electronic Imaging: Science & Technology, 1994.

[7] C. Lambrecht and O. Verscheure, “Perceptual Quality Measure using a Spatio-Temporal Model of the Human Visual System”, in Proc. of the SPIE, Vol. 2668, pp. 450-461, 1996.

[8] A.B. Watson, J. Hu, and J.F. McGowan “Digital video quality metric based on human vision” in *Journal of Electronic Imaging 10*(1), pp. 20-29, 2001.

[9] J.Malo, A.M. Pons, and J.M. Artigas, “Subjective image fidelity metric based on bit allocation of the human visual system in the DCT domain”, Image and Vision Computing, Elsevier, n. 15, pp. 535-548, 1997.

[10] M. Masry, S.S Hemami, Y. Sermadevi, “A Scalable Wavelet-Based Video Distortion Metric and Applications”, IEEE Trans. On Circuits and Systems for Video Technology, Volume 16, Issue 2, pp. 260-273, 2006.

[11] Z. Wang, A. C. Bovik, L. Lu and J. Kouloheris, "Foveated wavelet image quality index," SPIE’s 46th Annual Meeting, Proc. SPIE, Application of digital image processing XXIV, vol. 4472, July-Aug. 2001.

[12] Z. Wang, A.C. Bovik, “A Universal Quality Index” IEEE Signal Processing Letters, vol. 9, no. 3, pp. 81-84, March 2002

[13] Z. Wang, A.C. Bovi, H.R. Sheikh, E.P. Simoncelli “Image Quality Assessment: From Error Visibility to Structural Similarity” IEEE Transactions on Image Processing, vol. 13 no. 4 April 2004

[14] Z. Wang, E.P. Simoncelli “An adaptative linear system framework for image distortion analysis” Proc. IEEE Inter. Conf. Image Pro. Genoa, Italy Sep 2005

[15] Z. Wang, E.P. Simoncelli “Traslation insensitive image similarity in complex wavelet domain” Proc. IEEE Inter. Conf. Acoustic, Speech & Signal Processing Vol II Pages 573-576, March 2005

[16] Z. Wang, L. Lu, A.C.Bovik “Video quality assessment using structural distortion measurement”, IEEE International Conference on Image Processing, Sept. 2002.

[17] Z.Wang, A.C.Bovik, B.L.Evans “Blind Measurement of Blocking Artifacts in Images” Proc. IEEE Int. Conf. on Image Processing, Sep. 10-13, 2000, vol. III, pp. 981-984, Vancouver, Canada.

[18] E.P. Simoncelli “Modeling the joint statistics of images in the wavelet domain” Proc. SPIE 44th Annual Meeting, vol.3813, pp.188-195, Denver,Colorado. Jul. 1999.

[19] H.R. Sheikh, A.C. Bovik, L.Cormack “No-reference quality assessment using natural scene statistics: jpeg2000”, IEEE Trans Image Process. 2005 Nov;14(11):1918-27.

[20] Z. Wang, E.P. Simoncelli “Reduced-reference image quality assessment using a wavelet-domain natural image statistics model” Human Vision and Electronic Imaging X, Proc. SPIE, vol. 5666. 2005

[21] H.R. Sheikh, A.C. Bovik, G. Veciana, “An information fidelity criterion for image quality assessment using natural scene statistics” IEEE Transactions on image processing, vol 14, no. 12, Dec. 2005

[22] M.H. Pinson, S. Wolf “A new standardized method for objectively measuring video quality” IEEE Transactions on broadcasting, Vol. 50, No. 3. (Sep. 2004), pp. 312-322.

[23] S.Winkler, E.D. Gelasca, T.Ebrahimi “Perceptual Quality Assessment for video watermarking” itcc, p. 0090, International Conference on Information Technology: Coding and Computing, 2002.

[24] Z. Wang, H.R. Sheikh, A.C. Bovik “No-reference perceptual quality assessment of jpeg compressed images” IEEE International Conference on Image Processing, 477-480, Sep. 2002

[25] P.Marziliano, F. Dufaux, S.Winkler, T.Ebrahimi “Perceptual blur and ringing metrics: application to jpeg2000” Signal Processing: Image Communication, Vol. 19, No. 2. (February 2004), pp. 163-172.

[26] Z.Wang, A.C.Bovik, H.R. Sheikh, E.P. Simoncelli “Image Quality Assessment: From Error Visibility to Structural Similarity” IEEE Transactions on Image Processing, vol. 13, no.4, April 2004

[27] H.R. Sheikh, A.C. Bovik, "Image information and visual quality," IEEE Transactions on Image Processing, vol.15, no.2pp. 430- 444, Feb. 2006

[28] H.R. Sheikh, M.F. Sabir, A.C. Bovik, “A statistical evaluation of recent full reference image quality assessment algorithms” IEEE Trans.on Image Processing, Jan. 2006.

[29] P. Corriveau, et al., “Video quality experts group: Current results and future directions,” Proc. SPIE Visual Comm. and Image Processing, vol. 4067, June 2000.

[30] ISO/IEC 14496-10:2003. Coding of audiovisual objects part 10:advanced videocoding. ITUT Recommendation H264 Advanced video coding for generic audiovisual services, 2003.

[31] ISO/IEC 15444-1. Jpeg 2000 image coding system. Part 1:core coding system, 2000.

[32] J. Oliver, M. P. Malumbres, “Fast and efficient spatial scalable image compression using wavelet lower trees,” in Proc. IEEE Data Compression Conference, Snowbird, UT, March 2003.

[33] B. Watson,and L. Kreslake, “Measurement of visual impairment scales for digital video”, In Proc. SPIE Human Vision, Visual Processing, and Digital Display IX, vol. 4299, pp. 79-89, San Jose, CA, 2001.