No-Reference Metric for a Video Quality Control Loop

Jorge CAVIEDES
Philips Research, 345 Scarborough Rd, Briarcliff Manor
NY 10510, USA, jorge.caviedes@philips.com

and

Joel JUNG
Laboratoires d'Electronique Philips, 22 av. Descartes
Limeil-Brevannes, FRANCE, joel.jung@philips.com

ABSTRACT

In this paper we present an objective quality metric based on a combination of no-reference impairment metrics and image features that can be used for monitoring and quality control for in-service operation. We present the design principles for the quality metric, its performance, and discuss its applicability. The quality metric is strongly correlated with subjective test scores for a set of sequences including a variety of compressed and post-processed video contents. We also analyze the suitability of a local, real-time quality control loop based on our quality metric and local control modules such as post-processing, error concealment, and image enhancement.

Keywords: objective image quality, no-reference models, quality of service, video quality control, post-processing.

1. INTRODUCTION

Evaluation of video quality is generally performed using subjective tests, based on evaluation by naïve human subjects. These subjective methods are time consuming, and not applicable for in-service testing. Moreover, in-service applications do not have access to the original or reference video. Thus, objective, *full-reference* methods, i.e. methods based on comparison with the original video as reference, are not applicable for in-service testing either. However, methods that use either a reduced amount of information from the reference or no reference information at all are very well suited for in-service quality assessment.

If we start from the premise that an impairment-free image is a high quality image, then a *no-reference* method based on impairment metrics would be suitable for in-service quality monitoring. Also, if the metric complexity allows real time software or hardware implementation, it would also be suitable for the development of a continuous quality control system.

A real-time quality control system is the key enabler of any quality of service model for real-time, interactive, video services over any type of communication network. Feedback provided by such an objective evaluation method could be helpful to rapidly adapt the processing to the current quality of transmission, or to monitor the quality of service. The potential returns are high especially if standards are adopted to assess objective quality at all points of interest along the video chain including broadcasting and reception stages.

Published research on quality assessment, based on impairment metrics, has shown promising results, particularly for methods based on blocking artifact metrics [4,5,9,12]. Other no-reference approaches are based on intrinsic image features [11] or perceptual quality metrics [1,3,6,10].

In previous work we have underlined the relevance and potential of impairment metrics for no-reference quality metrics [2]. Based on that work, we are pursuing the development of an overall quality measure (OQM) for MPEG coded video with particular emphasis on minimum set of metrics, perceptual and mathematical properties of the metrics, appropriate reference subjective testing, and design of a quality control loop.

In this paper, Section 2 deals with subjective testing issues relevant to the creation of objective quality metrics. In Section 3 we present the properties of the impairment metrics, and in Section 4 the quality metric. Section 5 shows the results obtained so far with the OQM. Section 6 addresses the basics of a quality control loop design; and finally, section 7 presents a discussion of the OQM approach.

2. SUBJECTIVE TESTING

Subjective testing provides the reference scores (rank order, mean scores, and variance) necessary to validate objective quality metrics. The goal of an objective quality metric is to predict the same rank order, and to show a high correlation with the mean subjective scores.

The main problem of subjective testing is the score variability. Variability appears to depend on the rate at which subjects learn what to look for in the image, individual preferences, and individual sensitivity to the defects observed. Other variability sources include affinity with the content and focus of attention.

One way to eliminate variability is to use a collective judgement by a panel of top experts, and use it as the unique subjective reference score. However, pure subjective testing is aimed at finding the judgement of an average subject.

At this stage in our research we have chosen to work with a combination of evaluation by advanced experts and subjects with expertise in video processing as a way to reduce variability without loosing entirely the probabilistic nature of subjective scores.

2.1. Test set description

The test has been built taking into account the following requirements:

- Content criticality,
- Multiple compression rates,
- Possible improvement of quality along the video chain, due to post-processing,
- Individual effects of different impairments on quality,
- Broad quality range.

Eight full-resolution sequences of varied content and characteristics have been MPEG-2 coded at bit-rates between 1.5 Mbit/s and 10 Mbit/s. A proprietary post-processing algorithm has been applied to these sequences. The test set also includes sequences affected by one impairment at a time (i.e. blocking, ringing, corner outlier).

2.2. Subjective test procedures

We ran subjective tests in order to establish the rank order of the test sequences using a quality scale from 1-10, later normalized to 1-100.

We have also done the following tests:

1. Score a set of sequences affected each by one impairment only, in order to find those that lie one JND (Just Noticeable Difference [7]) away from the original sequence (called 1-JND test). This test is the means to calibrate individual impairment metric according to the subjective quality scale.
2. Find the rank order of the sequences in the 1-JND set in order to identify any perceptual imbalances among impairments (called 1-JND set-ranking test).

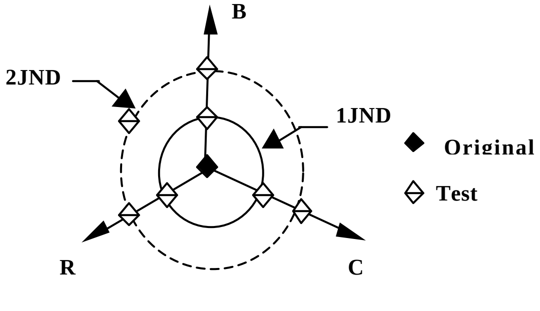

Figure 1. OQM space and test points.

The 1-JND test consists in subjectively scoring sequences in which one impairment has been increased while the others remain approximately constant (see in Figure 1 the case of three impairment metric axes B, R, and C). The impairment under test is increased to a point where subjects can notice a difference. The levels at which subjects notice a difference are called perception factors JND_B, JND_R, and JND_C. This test is way to introduce multi-dimensionality to the subjective quality assessments, establish anchor points on the critical 1-JND loss envelope, and carry on the calibration into the overall quality metric.

The 1-JND set-ranking test consists of a two-alternative forced choice (2AFC) method on all possible pairs within all the sequence set 1-JND away from the original.

3. PROPERTIES OF IMPAIRMENT METRICS

3.1. Mathematical properties

In order to build our quality metric we must choose impairment metrics considering:

- Linear independence,
- Perceptual impact of each impairment,
- High precision with a minimum of false results,
- Selection of a set of metrics complete enough to account for most relevant impairments.

We have chosen five impairment metrics which measure blocking, ringing, corner outliers, noise, and blur. These impairments appear to be uncorrelated because they arise from different types of image content.

Smooth regions are affected by blocking, strong natural edges are affected by ringing, noise affects uniform areas, and blur depends on contrast and sharpness of edges and textures. In our experience, the measures of the impairments do not show any systematic correlation in a variety of video sequences coded at several different bitrates.

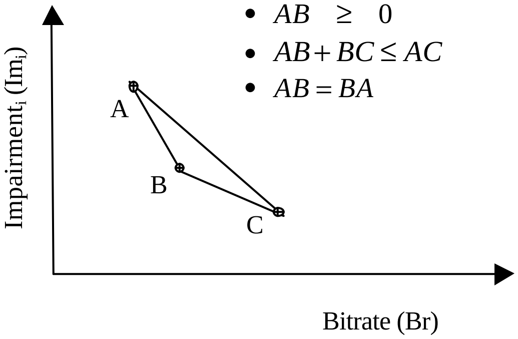

Figure 2. Properties of impairment metrics.

An impairment metric, and a combined metric, must be above all a true mathematical metric. The properties of a mathematical metric are:

1. The distance between two points is always greater or equal to zero; and the distance is zero if and only if the points are identical.

2. The distance between two points is symmetric.
3. The triangular inequality holds for any three points (see in Figure 2 an impairment metric plotted as a function of compression).

Impairment metrics are deterministic and generally monotonic functions of compression.

3.2. Quality units

In contrast with the mathematical metric, the subjective perception of impairments has its own scale, and there is a variance associated with the values.

The scale problem requires a transformation from impairment scale to perceived quality scale. The variance can be dealt with by using a resolution larger than the variance. The challenge is to still have a metric after these transformations.

In order to build an overall quality metric that expresses measures in quality units, these are the issues that must be solved:

1. A practical quality scale and its dynamic range (i.e. the most commonly used and most sensitive region of the scale) must be developed from the impairment metrics. The behavior of the final objective metric must correlate with subjective test scores, and show the mathematical properties explained above particularly within the dynamic range of the quality scale.
2. Noticeable changes in quality resulting from increasing or reducing individual artifact levels must be reflected correctly by the overall metric.
3. The resolution of the quality scale for each metric must be at least that of the subjective test scores.

4. THE QUALITY METRIC

In this section, we explain the working principles of each metric, performance, quality scale, and dynamic range. Metric performance is assessed in terms of its ability to make a maximum detection of real impairments, and a minimum number of errors.

4.1. Blocking artifact

Blocking artifacts are the best known MPEG artifacts. They show as discontinuities at the borders of 8x8 blocks as a consequence of strong, independent quantization of block DCT coefficients.

We have selected a blocking artifact metric developed in our laboratory based on its ability to detect visible discontinuities at block edges with a minimum of errors. The metric is representative of the amount of artifacts and their visibility. To verify *precision*, we highlight detected block edges on the actual image and then compare them against a sharpness-enhanced version of the same image to reveal any undetected artifacts. The amount of undetected artifacts was almost negligible in all tests. To verify *minimum error rate* we inspect visually that vertical and horizontal natural edges are not incorrectly detected as artifacts, and also test the detection algorithm on a varied set of non-coded for which the values given by the metric are always zero.

Regarding subjective quality scale, initial 1-JND tests indicate that the blocking artifact can introduce noticeable differences at a metric value of $JND_B = 60$. (When other impairments are low and approximately constant, i.e. selecting content that gets affected mainly by blocking). Above 1000 the image is so severely impaired that quality assessment is at the lowest level from that point on. Blocking is the best-behaved impairment. It increases monotonically with compression, but although one can find good examples of its consistent behavior. It is in general difficult to run experiments in which other impairments are constant.

4.2. Ringing artifact

Caused by coarse quantization of AC coefficients, ringing appears frequently near sharp edges that belong to low-activity regions of an image [12]. Depending on the orientation of the edges, the artifacts can appear as shimmering along the edges or multiple echoes (harmonics of the fundamental edge frequencies).

The principle of our ringing detection is based on searching for reduced strength edges (echoes) in the neighborhood of strong natural edges. The ringing metric is proportional to the number of pixels that belong to the false edges in an image. Verification of ringing metric precision and false detection rate can be done in an analogous way to the blocking metric verification explained before.

We observed that the ringing metric is not monotonic for low compression rates in at least one case. Further research on the implications of possible non-monotonicity of this metric must be done.

Initial tests on subjective perception were done by artificially creating sequences affected only by ringing. The results indicate that ringing becomes noticeable at about a level of 270.

4.3. Corner outlier artifact

Corner outliers are missing pixels that belong to strong contrast natural edges, i.e. they look conspicuously too light or too dark compared to their surroundings. The missing pixels are placed at the corners of 8x8 MPEG blocks [8]. A corner outlier is detected taking into account:

- Absolute luminance difference between the candidate corner pixel and the average of the group of four neighboring corner pixels to which it belongs,
- Perceptual visibility of that difference given the local average luminance,
- Probability that the candidate pixel is a natural pixel simply aligned with the grid. That is based on average number of corner outliers found if the grid is shifted to other possible locations on the image.

Verification of corner outlier metric performance is straightforward, as corner outliers are usually few and highly visible. In contrast with the two metrics explained before, the corner outlier metric is a probabilistic indicator of corner outlier level, and not based on a direct count of the actual pixels. The reason for this is mainly that our metric does not use coding information, i.e. quantization step size to detect corner outliers.

Nevertheless, each of the first three impairment metrics presented so far incorporates a simple implementation of Weber's law in order to account for perceptual visibility, plus empirical thresholds to separate natural image content from artifacts.

4.4. Noise, contrast, and sharpness

In addition to the impairment metrics presented above, we are also investigating noise, contrast and sharpness metrics. Their potential contribution must be assessed before we consider them for 1-JND testing. We have shown that these metrics significantly improve the performance of an overall quality metric based only on the three MPEG impairments discussed before.

4.5. The overall quality metric

In [2] we have proposed an Euclidean metric composed of the individual impairments. We now propose an Overall Quality Metric (OQM) that is already calibrated in JNDs as follows:

$$OQM = \sqrt{\left(B/JND_B\right)^2 + \left(R/JND_R\right)^2 + \left(C/JND_C\right)^2} \qquad (1)$$

where B is blocking, R is ringing, and C is corner outlier impairment; JND_B, JND_R, and JND_C are perception factors.

To assess the potential of noise, contrast and sharpness metrics, we have tested a weighted sum type of metric that includes them as well as blocking, ringing, and corner outliers (OQM-L). The results are presented in the following section.

5. PERFORMANCE OF THE OQM

The performance of the OQM, and OQM-L have been tested against subjective scores and also against full-reference objective metrics. The performance assessment takes into account ability to detect post-processing quality improvement, and correlation with expert ratings. Details are given in the following subsections.

5.1. Other quality metrics used for comparison

We have compared OQM against three full-reference quality metrics:

- The PSNR, calculated by averaging in time the PSNR of each frame.
- The IES-CPqD v.2.0, a full-reference quality evaluation method [13]. Scenes are segmented into flat, edge and texture regions and a set of objective parameters is computed separately within each of these contexts. The final predicted impairment level is achieved through a combination of these parameters.
- The PQR, the picture quality rating given by the Tektronix PQA200. PQR is based on a visual discrimination model that simulates the responses of human spatiotemporal visual mechanisms and the perceptual magnitudes of differences in mechanism outputs between source and processed sequences. From these differences, an overall metric of the discriminability of the two sequences is calculated.

5.2. Correlation with expert ratings

Figure 3 shows the ability of OQM to predict the subjective rank order. It shows for each sequence the correlation between the subjective score and OQM rating. In the best case, all sequences would be inside the enclosed area.

For OQM-L, the results are shown in Figure 4. The graph shows the 63 sequences in the test set ranked by quality from worse to best. For simplicity, in the Y-axis we show 5 minus the subjective score in order to have similar degradation curves. An overall correlation of 0.81 has been obtained for a non-optimized, initial formulation of OQM-L.

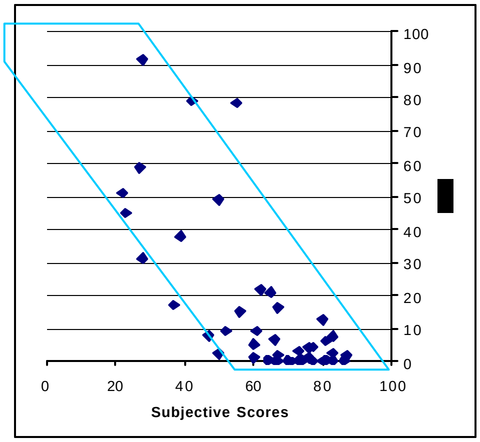

Figure 3. Rank order correlation.

As it could be expected, the quality of degraded sequences that show minimum blocking, ringing, and corner outliers is overestimated by OQM (seen as many points at or near the zero OQM level in Figure 3), while OQM-L shows a dramatically improved overall performance.

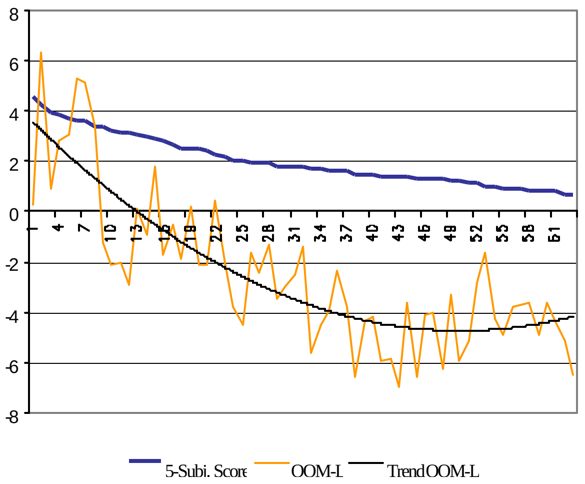

Figure 4. Performance of the quality metric including noise, contrast, and sharpness.

As a benchmark, we compared the 0.81 correlation of OQM-L against CPqD (see next subsection for performance of this and other metrics). The correlation obtained for CPqD was 0.87. Thus, the non-optimized performance for the no-reference OQM-L is encouraging at this stage. (Especially considering that the correlation for PSNR was 0.7 for the same test set)

5.3. Perception of post-processing improvements

Results given in Figure 5 show, for each method, how often the post-processed sequence has been declared of better quality than the MPEG2 decoded sequence.

As it can be observed, the improvement is best detected by the OQM method, overestimating in some cases the improvement perceived by the experts. CPqD also shows ability to detect post-processing improvement. In contrast, the improvement is not well detected by the PQR and the PSNR

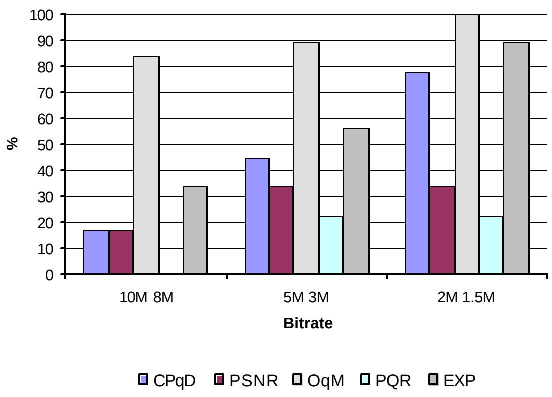

Figure 5. Detection of quality improvement for post-processed video sequences.

5.4. 1-JND set-ranking test

Results of the 1-JND set-ranking test confirm the results obtained by the 1-JND tests. The 2AFC method applied on all pairs of sequences that are 1-JND from the original show that 1-JND of any impairment is perceptually equivalent. People are unable to rank these sequences.

6. THE QUALITY CONTROL LOOP

In this section we discuss a local quality control loop acting on the video chain components starting from the decoder, including post-processing steps such as compression artifact reduction, and possibly transmission error concealment and image enhancement.

Figure 6 shows a section of the video chain with potential monitoring points, the quality profile as a function of distance from the source, and the control paths. The quality vs. distance-from-the-source graph indicates that when post-processing is involved, quality is not always lost, and thus quality improvements must be taken into account.

We use OQM as the monitoring signal that enables correction through local modules such as decoder and post-processor in order to improve quality in the presence of quality drops greater than 1JND.

OQM can also be used in a *reduced-reference* mode, i.e. comparing partial information among different monitoring points. If we consider a previous stage in the video chain a floating reference, we can compare impairment metrics and overall quality metric between the two points provided that the measures are time-stamped.

An example of a control scenario would be a decoder (or a transcoder at the output of a local storage system) set to work adaptively based on the quality at the output. In the event of a quality drop greater than 1-JND, the composition of the individual metrics, which are already calibrated in JNDs, allows acting upon the appropriate modules and choosing the best strategy to bring quality up to the desired level.

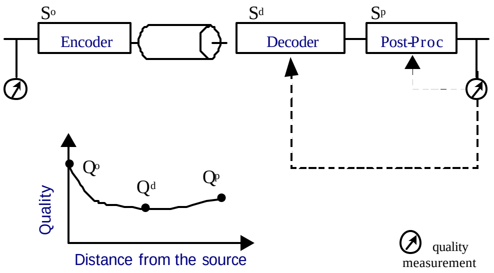

Figure 6. Quality control along the video chain.

7. DISCUSSION

We have presented a no-reference, overall quality metric based on impairment metrics and image features, and shown that it is possible to achieve a high correlation with subjective scores. We have addressed quality range and scale resolution issues and proposed testing procedures to achieve proper calibration of the individual and overall metrics in the critical region where the first JND loss takes place.

The initial results of a metric including blocking, ringing, corner outliers, noise, contrast and sharpness show significant correlation with subjective scores. This type of metric can overcome disadvantages of other models for objective quality, which give fidelity measures not suitable for measuring image improvement. We have also pointed out that unidimensional quality measures cannot be used in quality control systems to the extent that the components of the metric cannot be interpreted in order to develop a flexible, multilevel control strategy.

We have also proposed that quality metrics based on impairment metrics are suitable to develop no-reference and reduced-reference quality metrics. The main strength of our approach is the open design of the overall quality metric, which means that the impairment metrics can be replaced for new ones at any time if they show better performance or lower complexity.

Regarding quality control, we have touched upon a key issue: a mathematical metric is suitable for quality control, but it must also correlate with subjective test scores. More research must be done to refine the JND calibration of individual metrics and extend the range of such calibration.

Achieving high performance in the 1-JND region allows setting alarm levels for a quality control loop in both no-reference and reduced-reference modes. The feasibility of this approach depends on the computational complexity of the metric and its feasibility in real time systems. It also depends on the controllability and observability of the quality parameters as allowed by the architecture of the different modules on the video chain.

The issue of dynamic range of the quality scale needs further research. In principle the dynamic range of a quality metric is the union of the dynamic ranges of the base metrics. That dynamic range must be validated against the dynamic range of subjective scores.

ACKNOWLEDGEMENTS

The authors would like to acknowledge the CPqD Telecommunication Research and Development Center for providing the IESCPQDv2.0 binary for evaluation.

REFERENCES

[1] A.J. Ahumada, Jr. Et al., “Image quality: a Multidimensional Problem”, *in Digital Images and Human Vision,* ed. A.B. Watson, MIT Press, pp. 141-148, 1993.

[2] J.E. Caviedes, A. Drouot, A. Gesnot, L. Rouvellou, “Impairment Metric for Digital Video and their Role in Objective Quality Assessment”, *Visual Communications on Image Processing*, pp. 791-800, Perth, June 2000.

[3] M.P. Eckert and A.P. Bradley, “Perceptual quality metrics applied to still image compression”, *Signal Processing*, Elsevier, n. 70, pp. 177-200, 1998.

[4] C. Glassman, A. Peregoudov, A. Logunov, and V. Lichakov, “Video compression artefacts: predicting the perceptual ratings”, *Proceedings of IBC*, pp. 560-64, Amsterdam, 1999.

[5] S.A. Karunasekera and N.G. Kingsbury, “A distortion measure for blocking artifacts in images based on human visual sensitivity”, *IEEE Transactions on Image Processing*, vol. 4, n. 6, pp. 713-24, 1995.

[6] V. Kayargadde and J. Martens, “Perceptual characterization of images degraded by blur and noise: model”, *J. Opt. Soc. Am A*, vol. 13, n. 6, pp. 1178-1188, June 1996.

[7] J. Lubin and D. Fibush, “Sarnof JND vision model”, T1A1.5 Working Group Doc. No. 97-612, T1 Standards Committee, 1997.

[8] H.W. Park and Y.L. Lee, “A Postprocessing method for reducing quantization effects in low bit-rate moving picture coding”, *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 9, n. 1, pp. 161-71, 1999.

[9] M. Trauberg, “A new method of picture quality monitoring in MPEG-based systems”, *International Broadcast Engineer*, pp.39-45, November 1998.

[10] C.J. Van den Branden Lambrecht, O. Verscheure, “Perceptual Quality Measure Using a Spatio-Temporal Model of Human Visual System”, *Proceedings of SPIE,* vol. 2668, pp. 450-461, San Jose, January 1996.

[11] A.A. Webster, et al., “An objective video quality assessment system based on human perception”, *Proceedings of SPIE,* vol. 1913, pp. 15-26.

[12] H.R. Wu and M. Yuen, “Quantitative quality metrics for video coding blocking artifacts”, *Proceedings of Picture Coding Symposium 1*, pp. 23-26, Melbourne, Australia, 1996.

[13] ITU-T Study Group 12, Contribution COM12-39, “Video Quality Assessment using Objective Parameters based on Image Segmentation”, December 1997.