SUBJECTIVE AND OBJECTIVE QUALITY EVALUATION OF LAR CODED ART IMAGES

C. Strauss†, F. Pasteau†, F. Autrusseau‡, M. Babel†, L. Bédat†, O. Déforges†

† IETR UMR CNRS 6164 - Image Group
INSA of Rennes, France
‡ IRCCyN UMR CNRS 6597
Image and Videocommunications Team, France

ABSTRACT

Quality assessment is of major importance when designing and testing an image/video coding technique. Compression performances are usually evaluated by means of rate-distortion curves. However, the PSNR is commonly employed as the distortion measure. We hereby present a full quality assessment benchmark for the LAR (Locally Adaptive Resolution) coder. We conducted a subjective experiment, where nineteen observers were asked to assess the perceptual quality of LAR coded images under normalized viewing conditions. Furthermore, five objective quality assessment metrics were used in order to determine the most suitable metric for the LAR coder. Finally, both JPEG and JPEG200 images were generated and assessed during the subjective experiment in order to define the optimal quality metric which should be used when comparing the codecs' output images quality.

Index Terms— Image coding, Image segmentation, Quality assessment

1. INTRODUCTION

Digital museums aim at providing via the Internet digital versions of the original art items collected in a database on a server [1]. In doing so, museums tend to preserve their huge number of items and to widely spread associated cultural knowledge [2, 3]. Nevertheless communicating these materials over the Internet raises inherent security problems requiring hierarchical access policy [4]. In France, the C2RMF laboratory, connected to the Louvre museum, has digitized more than 300000 cultural items taken from French museums, in high resolution (up to 20000 × 30000 pixels). The resulting EROS database [5] is for the moment only accessible to art researchers whose work is directly connected with the C2RMF. The French TSAR project is designed to open the EROS database in a secure, efficient and user-friendly way that involves cryptography and watermarking as well as compression and region-level representation abilities.

The LAR codec addresses the last two main objectives, namely a scalable compression scheme efficient from low-bit rates up to lossless coding together with a free hierarchical region representation. This region representation enables chrominance coding at region level leading to both interesting compression ratio and functionalities. This paper is especially focused on the quality evaluation of encoded color images.For this purpose, different evaluation methods will be used. The PSNR is widely used to evaluate the perceived quality of compressed images, but its performances are arguable. Evidently a subjective quality experiment is the most efficient way to evaluate the performances of a compression scheme. During a subjective experiment, observers are enrolled and asked to assess the quality of distorted images on a predefined distortion scale. However, as we will see in Section 3, subjective experiments are very restrictive and time consuming. Furthermore, very few image processing labs have access to the experimental setup needed for such experiments (room under normalized illumination, normalized image background using a characterized monitor). Thus, objective quality metrics are commonly used to assess the rate-distortion performance of the coding technique. However, the objective metrics performances strongly fluctuates depending on the studied coding algorithm. Effectively, each lossy compression algorithm introduces very specific visual distortions, such as blocking, ringing or blurring artifacts, consequently, the selection of the most appropriate objective quality metric for a given compression technique is critical, i.e. the metric providing the best correlation with human judgement. Therefore, our aim in this work is to determine the best objective quality metric for LAR coded images. Effectively, determining the highest performances quality metric for the LAR will be of great help in future developments to improve the LAR codec.

This paper is organized as follows. Section 2 introduces the LAR coding method devoted to low-bit rate region-based color encoding. Section 3 presents both the subjective experiment protocol and the objective quality metrics used. Experimental results are given in Section 4, were we will see that two out of the five tested metrics showed better overall performances. Finally, Section 5 concludes our experiments.

2. LOW BIT-RATE REGION-BASED COLOR LAR CODEC

The LAR (Locally Adaptive Resolution) codec relies on a two-layer system. The first layer, called Flat coder, leads to construct a low bitrate version of the image with good visual properties. The second layer deals with the texture that is encoded through a spectral coder. This image decomposition into two sets of data is performed conditionally to a specific quadtree representation. From this, an original segmentation process can be conducted at both the coder and the decoder [6]. In the following, we describe the major features of a scalable version LAR coder (S+P) [7] associated to the dedicated region representation.

This work is supported by the French National Research Agency as part of the TSAR project. All art pictures are the property of the C2RMF laboratory. F.A. whishes to thank R. Pepion & R. Cousseau for their valuable input running the subjective experiment

2.1. Scalable LAR Coder

The basic idea is that local resolution, in other words pixel size, depends on local activity. This leads to the construction of a variable resolution image based on a quadtree data structure, encoded in the Flat coding stage. Thanks to this type of block decomposition, their size implicitly gives the nature of the said block. In a lossy context, this image content information controls a quantization of the luminance where large blocks require fine quantization (in uniform areas, human vision is strongly sensitive to brightness variations) while coarse quantization (low sensitivity) is sufficient for small blocks. To fit the Quadtree partition, dyadic decomposition is carried out. The first and second layers are processed through two successive partial pyramidal decomposition. The image representation content is preserved: the first decomposition reconstructs the low-resolution image (LAR-image) while the second one processes the local texture information. Thus, the first pyramid pass performs a conditional decomposition in accordance with the Quadtree partition (Fig. 1).

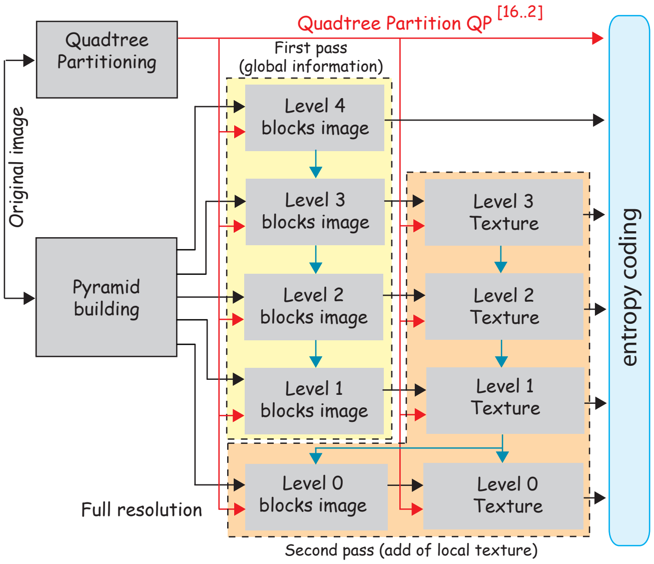

Fig. 1. Interleaved S+P LAR luminance coder.

2.2. Low coding rate region representation

Even if efficient context-based methods adapted to Quadtree based region partition compression have been already developed, prohibitive partition coding cost stays one of the principal restrictions to the evolution of content-based coding solutions. Dedicated to the LAR, the segmentation proposed in [6] is an efficient adaptation of the split/merge methods that tackles coding constraints. Given that both splitting process and luminance block image encoding have been realized by the flat LAR, merging process only deals with the finest partition i.e. the Y-block image (Cr/Cb block-images are not first considered). To take advantage of color information, a ”chromatic control” principle is defined and included in the merging process previously described. This chromatic control generates binary information for each luminance-based merging attempt to control the merging process (Figure 2).

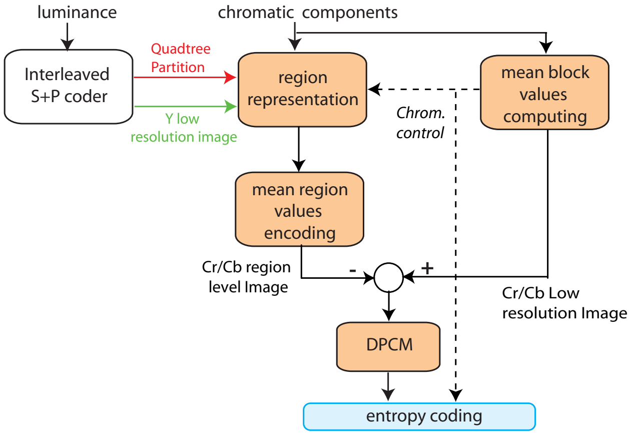

Fig. 2. Region based chromatic components coding.

2.3. Chromatic components coding

To take advantage of the previous segmentation process, chromatic components can be efficiently encoded at region level, requiring naturally one value per region. This compact representation leads to very low bit rate encoding, where color components are compliant with the image content. An enhancement layer can be obtained on chromatic data when designing a block-level predictive encoder that takes into account previous region values.

3. QUALITY ASSESSMENT

The best way to assess the perceived quality of coded images is evidently to run a subjective experiment, where human observers are asked to assess the quality of distorted images according to standardized viewing protocols. Unfortunately, such experiments are very restrictive, time consuming and require a very specific setup (normalized viewing conditions according to ITU recommendations).

3.1. Subjective Quality Assessment

In this work during the subjective experiment both the original and coded images were simultaneously presented to the observers on the viewing display. A CRT monitor was used and the viewing distance was set to four times the image height. The room background luminance was set accordingly to the ITU recommendation ITU-R-BT.500-1. After a viewing display of 8 seconds, the observers had to rate the quality of the presented coded image regarding the original image, which was explicitly known (always displayed on the left hand side). Eight original images were used, four were setected from the Microsoft JPEG database (*P02*, *P06*, *P09* and *P26*), and four were art images (scanned paintings and sculptures photographs) from the C2RMF EROS database[1] . These latter are represented on Fig. 3 [2]. For every original image, three encoders were considered (JPG, J2K, LAR), at five encoding rates. Once the observers scores collected, an average is computed for every image and corresponding distortion to produce the Mean Opinion Score (*MOS*). The quality scale is depicted on Table 1. Each session was about 25 to

[1] Microsoft DB: sftp://etro6.vub.ac.be, C2RMF DB: http://www.c2rmf.fr/

[2] Images references: (a) Notre-Dame de Grasse, Toulouse, FZ29646-ds11559, (b) Le scribe, Louvre, FZ22535-a10036, (c) and (d) Vieillard et enfant, Ghirlandajo, Louvre, F2880-dv57503

Fig. 3. The art images used in this experiment.

30 minutes long, depending on the time needed by the observers to assign a score.

5	Imperceptible
4	Perceptible, but not annoying
3	Slightly annoying
2	Annoying
1	Very annoying

Table 1. The used quality assessment scale

3.2. Objective Quality Assessment

As previously explained, subjective experiments are restrictive, as a very specific experimental setup is needed, and thus, efficient objective quality metrics are highly desirable for a fast and easy prediction of the observers scores. Several works have recently focused on the design of such quality metrics. In the following, five quality metrics are used (C4[8], VSNR[9], VIF[10], SSIM[11], WPSNR-PIX as referred in ITU-R-BT.601) to assess the quality of LAR coded images at different compression ratios. The performances of each metric are evaluated using the rank correlation, linear correlation, RMSE, weighted RMSE, Outlier Ratio (consistency) and Kappa coefficient (measuring the agreement). Interested readers should refer to the VQEG meeting reports[3] for further details on these measures. Both C4 and VSNR are based on advanced Human Visual System (HVS) features, C4 includes a multichannel model operating a Fourier subband decomposition. VIF mixes both Natural Scene Statistics and some basic properties of the HVS. And finally, SSIM uses simple statistics on the images. A mapping function was used on the metrics outputs, as recommended by the VQEG Multimedia TEST PLAN. This mapping function allows to rescale the objective measures to fit in the range of the *MOS*, which in the presented experiment is [1,5], as shown in Table 1. Once the mapping function computed for every quality metric, the predicted *MOS* (often referred to as the *MOSp*) may then be compared to the observers *MOS*.

4. EXPERIMENTAL RESULTS

As previously explained in sections 3.1 and 3.2, a subjective experiment was designed and the scores from 19 observers with normal or corrected human vision were collected. Furthermore, the main contribution of this work is to determine the most suitable objective quality metric for the LAR coding technique. Thus, five objective quality metrics were evaluated. We present in this section the experimental results in terms of metrics performances regarding the subjective scores. Figure 4 presents for all metrics, the *MOS* plotted as a function of the metrics' *MOSp*. Usually, better performances are obtained when the symbols follow a $y = x$ line, which seems to be particularly true for C4 and VIF. Furthermore, Fig. 5 represents the distortion as a function of the bit-rate for all tested quality metrics. This plot confirms the better performances of both C4 and VIF, for which the curves are very close to the *MOS*.

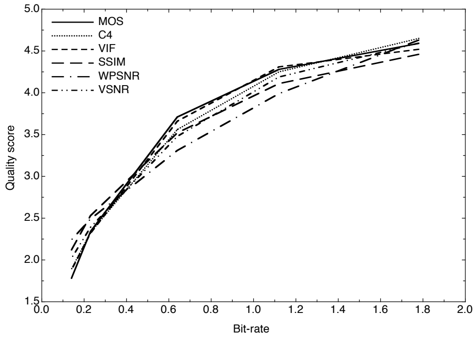

Fig. 5. Rate distortion curves for the five metrics

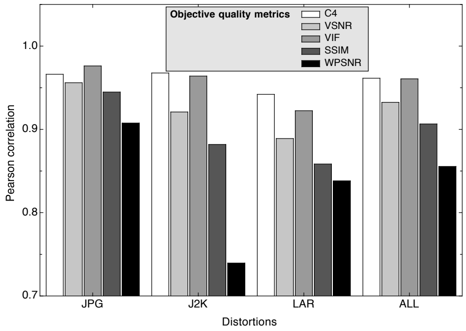

Fig. 6. Pearson correlation between quality metrics and human judgement respectively for JPG, JPK, LAR and the whole database

Figure 6 shows for all tested metrics the Pearson correlation between the *MOSp* and the *MOS*. On the x-axis, the 40 distorted images successively for JPEG, JPEG2000, LAR coding and for the whole database including all codecs (120 images). It clearly appears on this plot that when comparing quality measures between JPEG, JPEG2000 or LAR coding, either C4 or VIF should be preferred. The metrics' performances are depicted on Table 2 for the LAR codec, where both the linear and rank correlation are given along with the RMSE, weighted RMSE (WRMSE), Outlier Ratio and Kappa coefficient for every metric. It appeared that for most performances measures, the C4 and VIF metrics outperformed the other metrics. WPSNR, which is commonly used for rate-distortion evaluation of coding techniques, showed the worst correlation with the human judgement. Briefly, the highest are the correlations, the

[3] http://www.its.bldrdoc.gov

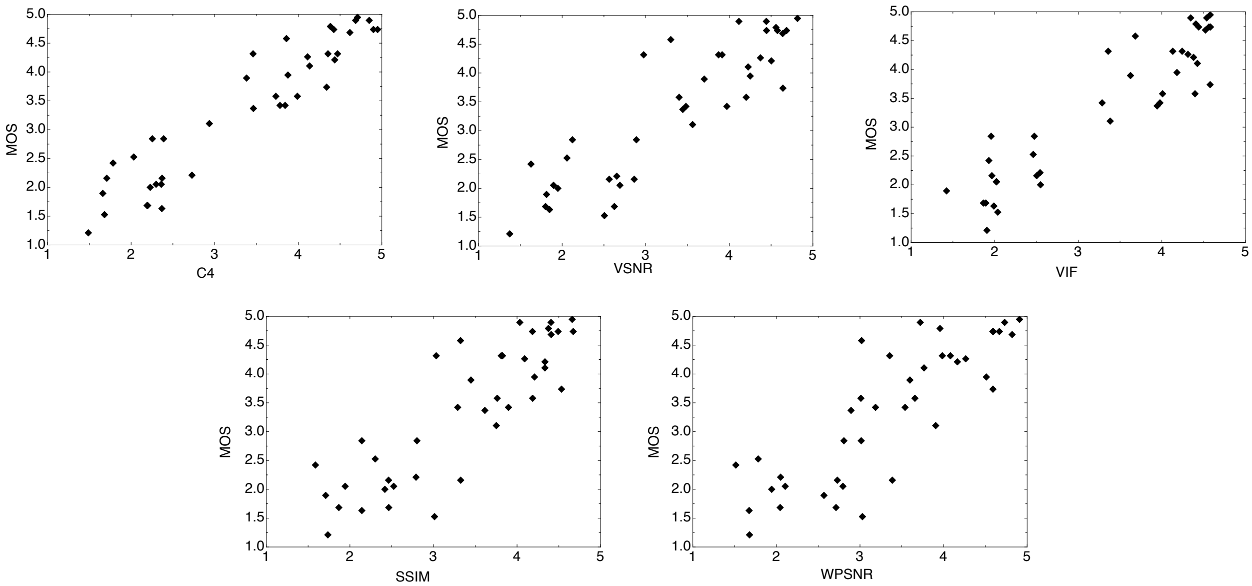

Fig. 4. *MOS* versus predicted *MOS* for the five tested quality metrics.

lowest are the RMSE and Outlier Ratio, the best are the metric's performances. Furthermore, VQEG has set a lower Kappa threshold at 0.4, below which, metrics are considered as inefficient.

	C4	VSNR	VIF	SSIM	WPSNR
WRMSE	1.267	1.777	1.563	2.054	2.248
RMSE	0.390	0.533	0.450	0.599	0.635
RankCorr	0.916	0.860	0.875	0.828	0.831
OutlierRatio	0.125	0.200	0.275	0.350	0.300
Kappa	0.629	0.411	0.464	0.328	0.442
LinCorr	0.942	0.889	0.922	0.858	0.838

Table 2. Performances of the five quality metrics for the LAR coding

5. CONCLUSION

A subjective quality experiment was conducted specifically for LAR coded images on a set of eight input images, each being distorted at five compression rates. Nineteen observers with correct vision were enrolled and had to rate the quality of the distorted images on a [1,5] quality scale. Furthermore, five quality metrics were tested on the so obtained 40 distorted images. Six performance measures were used in order to rank the objective metrics altogether. We concluded that the metrics C4 and VIF were the most appropriate when assessing the quality of LAR coded images. Besides LAR coded images, a subjective experiment was conducted on both JPEG and JPEG2000 images. Our goal was to determine which metric should be used when comparing the coded images quality. The C4 and VIF metrics presented the best performances for all kind of coding distortions.

6. REFERENCES

[1] G. F. MacDonald, "Digital Visionary," *Museum News*, March/April 2000.

[2] X. Chen, H. Ou, X. Luo, M. Chen, Y. Zhang, K. Hao, and S. Mi, "The Progress of University Digital Museum Grid," in *Proc. IEEE International Conference on Grid and Cooperative Computing Workshops, GCCW'06*, 2006.

[3] P. F. Marty, "The Digital Museum in the Life of the User," in *Digital Humanities*, October 2005.

[4] M. S. Shapiro, "Managing Museum Digital Assets: A Resource Guide for Museums," Tech. Rep., International Intellectual Property Institute, 2001.

[5] D. Pitzalis, R. Pillay, and C. Lahanier, "A new Concept in high Resolution Internet Image Browsing," in *10th International Conference on Electronic Publishing (ELPUB)*, June 2006.

[6] O. Déforges, M. Babel, L. Bédat, and J. Ronsin, " Color LAR Codec: A Color Image Representation and Compression Scheme Based on Local Resolution Adjustment and Self-Extracting Region Representation," *IEEE Trans. on Circuits and Systems for Video Technology*, vol. 17, no. 8, pp. 974–987, August 2007.

[7] M. Babel, O. Déforges, and J. Ronsin, "Interleaved S+P Pyramidal Decomposition with Refined Prediction Model," in *IEEE International Conference on Image Processing, ICIP'05*, Genova,Italy, Septembre 2005, vol. 2, pp. 750–753.

[8] M. Carnec, P. Le Callet, and D. Barba, "Objective quality assessment of color images based on a generic perceptual reduced reference," *Signal Processing: Image Communication*, vol. 23, no. 4, pp. 239–256, April 2008.

[9] D. M. Chandler and S. S. Hemami, "Vsnr: A wavelet-based visual signal-to-noise ratio for natural images," *IEEE Trans. on Image Processing*, vol. 16, no. 9, pp. 2284–2298, 2007.

[10] H.R. Sheikh and A.C. Bovik, "Image information and visual quality," in *IEEE Trans. on Image Processing*, May 2005.

[11] Z. Wang, A.C. Bovik, H.R. Sheikh, and E.P. Simoncelli, "Image quality assessment: From error visibility to structural similarity," *IEEE Transactions on Image Processing*, vol. 13, no. 4, pp. 600–612, 2004.