USING H.264/AVC-INTRA FOR SEGMENTATION-DRIVEN COMPOUND DOCUMENT CODING

Alexandre Zaghetto, Ricardo L. de Queiroz

Department of Electrical Engineering, Universidade de Brasilia, Brazil
zaghetto@image.unb.br, queiroz@ieee.org

ABSTRACT

In this paper we explore H.264/AVC operating in intraframe mode to compress a mixed image, i.e. composed of text, graphics and pictures. Even though mixed contents (compound) documents usually require the use of multiple compressors, we apply a single compressor for both text and pictures. For that, distortion is taken into account differently between text and picture regions. Our approach is to use a segmentation-driven adaptation strategy to change the H.264/AVC quantization parameter on a macroblock by macroblock basis, i.e. we deviate bits from pictorial regions to text in order to keep text edges sharp. We show results of a segmentation driven quantizer adaptation method applied to compress documents. Our reconstructed images have better text sharpness compared to straight unadapted coding, at negligible visual losses on pictorial regions. Our results also highlight the fact that H.264/AVC-INTRA outperforms coders such as JPEG-2000 as a single coder for compound images.

Index Terms— Compound document coding, segmentation-driven image coding, H.264/AVC

1. INTRODUCTION

The newest video coding standard, the H.264/AVC [1], has been well explained in the literature [2]-[5]. Many papers have illustrated its performance showing many comparative results against coders such as MPEG-2. Apart from the factor-of-two improvement over other standards, there are a few unexpected advantages that come with the AVC package.

H.264/AVC is a video compression standard and it was not conceived to be applied as a still image compression tool. Nevertheless, the many coding advances brought into H.264/AVC, not only set a new benchmark for video compression, but they also make it a formidable compressor for still images [6], [7]. One of the components of these advances is the intraframe macroblock prediction method, which, combined with the context-adaptive binary arithmetic coding (CABAC), turns the H.264/AVC into a powerful still image compression engine. If we set our H.264/AVC implementation to work on a sole "INTRA" frame it will behave as a still image compressor. We refer to this coder as AVC-I. The big surprise is that it also outperforms previous state-of-art coders such as JPEG-2000 [8].

Gains of the AVC-I over JPEG-2000 are typically in the order of 0.25dB to 0.5dB in PSNR for pictorial images [7], [9]. However, for compound images (mixed pictures and text) the PSNR gains are more substantial, even surpassing the mark of 3 dB improvement in some cases, as shown in Fig. 1.

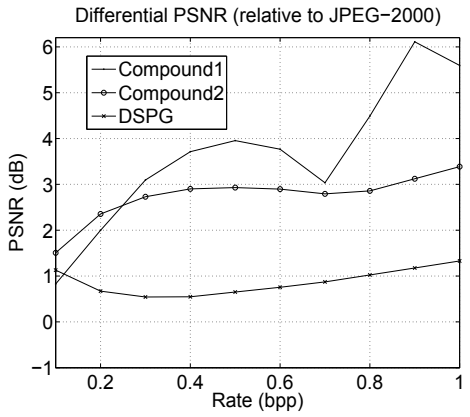


Fig. 1. Differential PSNR (relative to JPEG-2000) plots comparing AVC-I against JPEG-2000 for "compound1", "compound2", and "DSPG" images. The compoundN images belong to the JPEG-2000 test set. Because of the very large size of "compound2" we selected only a portion of it for tests.

Electronic documents are basically represented in two forms: vectorial or raster [10]. It is not much of a challenge to compress vectorized documents since each object can be compressed individually and the whole file can be compressed losslessly. The real challenge is to compress rasterized documents. Compound documents are assumed here as raster documents which contain a mix of text and pictorial contents.

Compression algorithms are developed with a particular image type, characteristic and application in mind and no single algorithm is best across all types of images or applications. When compressing text, it is important to preserve the edges and shapes of characters accurately to facilitate read-

This work was supported by a grant from HP Brasil.

ing. The human visual system (HVS), however, works differently for typical continuous-tone images, better masking high-frequency errors [11].

Compound raster documents have typically been compressed as a single image. However, different compression algorithms may be applied to each of the regions of the document. That is the way multiple-coder based algorithms work [12]. Instead of a multiple-code approach, this paper proposes a single-coder algorithm based on a modified version of the AVC-I that adjusts itself as an effort to encode text and pictorial regions differently.

It is important to place our coder within the proper application context. We are concerned with a distributed scenario such as the web, where a document is stored once, somewhere, and many users are to retrieve and decode the document at a later time. In this scenario, the encoder has no complexity restriction.

2. SEGMENTATION-DRIVEN RATE ALLOCATION

A few authors dealt with compressing documents with one coder. For example, Konstantinides and Tretter [13] used adaptive quantization within the JPEG extensions framework to compress compound (mixed) images. The idea is to use less aggressive quantizer steps for text regions in order to keep edges sharp, while being more forgiving to high frequency losses in pictures. Ramos and De Queiroz [14] used a single JPEG coder for the compression of mixed documents, stealing bits from background and images to give to text and sharp graphics edges. For that, thresholding was used, which is a technique that analyzes each quantized DCT coefficients in JPEG, computing how many bits it costs to be transmitted and how much distortion it would cause to the reconstructed image if it is not transmitted at all. In effect, thresholding is an optimization technique for JPEG based on the rate-distortion analysis of every block. Coefficients which are not rate-distortion efficient are discarded (thresholded). It was just a case of making the rate-distortion analysis adaptive depending on the image contents.

In general, for RD optimized transform coding, the signal is divided into units x_i, each contributing to the overall bit-rate R by R_i bits, i.e. $R = \sum_i R_i$. Distortion is some function of the quantization error $\hat{x}_i - x_i$, where $\hat{x}_i$ is the reconstructed unit. The global distortion is given by

$$D = f(\{\hat{x}_i - x_i, \forall i\}) \quad \left(\text{e.g.} = \sum_i ||\hat{x}_i - x_i||\right) \quad (1)$$

By using a well behaved distortion function such as MSE, then $D = \sum_i D_i$ where D_i is the distortion for the i-th unit as $D_i = ||\hat{x}_i - x_i||$. RD optimization involves the minimization of a cost function $J = R + \lambda D$, where λ is a Lagrangian multiplier. Hence,

$$J = \sum_i R_i + \lambda ||\hat{x}_i - x_i|| \quad (2)$$

We imply a space varying meaning for distortion as opposed to adapting the algorithm, i.e. $D_i = ||\hat{x}_i - x_i||u_i$, where u_i is a distortion weighting factor specific for the i-th unit. In conventional human visual system weighted error measures, we can use a frequency-based weighting system in the transform domain. Since the HVS response is not completely understood and cannot be easily modeled, one can classify the image blocks into a discrete number of representative classes and devise HVS weights for each of the classes. For simplicity we assign weights u_i for the error norm rather than weights in the transform domain. Hence,

$$J = \sum_i R_i + \lambda_i ||\hat{x}_i - x_i|| \quad (3)$$

where $\lambda_i = u_i \lambda$.

In H.264/AVC, there are many decisions per block or macroblock that the encoder has to make. Examples are the division of the macroblock for motion estimation, the transform to be used (4×4 or 8×8), the intraframe block prediction mode, etc. These decisions are all made by rate-distortion considerations. Apart from that we can always use thresholding based on RD too, just like in JPEG. In all these cases we have to make an RD analysis per macroblock.

H.264/AVC allows for the change of the quantizer parameter Q_p at each macroblock. The adjustment of λ, or λ_i, in the quantization step, is translated into an adjustment of Q_p by an exponential equation. The quantizer adjustment is the most effective way to control rate and distortion. It controls more intensively the RD balance than for example using RD analysis to select the best macroblock prediction mode, or the size of the DCT. Therefore we can cut corners and adjust RD (minimize J) by modifying directly the quantizer parameter at each macroblock.

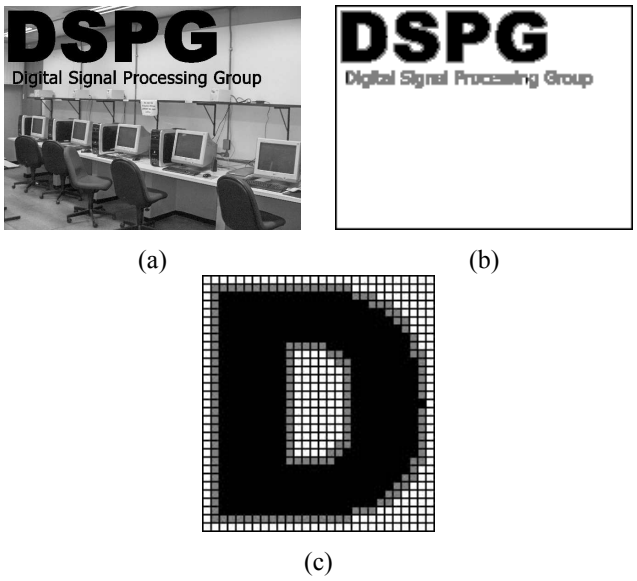
We propose to adapt the analysis on a macroblock by macroblock basis to be more economic in some blocks as opposed to others. First, we apply a region classification algorithm that will identify text and pictorial regions. This classification algorithm is derived from an edge detector and needs to identify edges belonging to text as opposed to textures. We assume that in these text regions the viewer would pay greater attention to edges. In our tests, we use Fan's text segmentation algorithm [15].

The next step is to classify each macroblock (16x16 pixels block), denoted here as MB. The binary image containing the segmented text is analyzed and each MB is classified as type 0, 1 or 2 and a coding mask is constructed. MBs of class 0 (pictorial regions) are composed exclusively by pixels marked as background. Class 1 MBs (text interior regions) are those composed exclusively by pixels marked as text. MBs which present a mixture of background and text interior, in any proportion, are considered as class 2 (text border

MBs). Fig. 2(b) shows the coding mask for the image shown in Fig. 2(a). Fig. 2(c) shows an enlarged portion of "D" in "DSPG". To make it easier to visualize, MB classes 0, 1 and 2 were represented as white, black and gray, respectively.

The coding mask is passed on to a modified version of AVC-I, which will adapt the value of Q_p for each MB, according to the class it belongs. The idea is to "transfer" quality of a MB class to another. Class 0 and 1 regions are encoded with a quantizer parameter Q_p, while class 2 regions are encoded with a quantizer parameter Q_{pText}, being $Q_{pText} < Q_p$. This means that more compression is applied where there is texture, and less compression is applied to the text letter borders. This algorithm is referred to as H.264/AVC-INTRA Compound, or simply AVC-C.

Alternatively, it can be used to encode a region of interest (ROI) differently from the rest. The idea is to allow a single H.264 coder to compress more efficiently compound images.



(a) (b)

(c)

Fig. 2. Segmentation/classification algorithm: (a) original grayscale image (2592x1952 pels); (b) its coding mask; (c) zoom around character "D" in "DSPG". The grid shows the 16x16 pixels macroblocks.

3. THE TEXT VS. PICTURE BALANCE

We want to lower the quality of pictorial and text interior regions to improve text border regions until they become sufficiently clear, without compromising the quality of the whole document. Our Q_p and Q_{pText} selection algorithm works as follows:

1) A bitrate R is chosen.
2) A bitrate variation δr around R is set.
3) Among all possible (Q_p, Q_{pText}) combinations, those which present bitrates inside the interval $R \pm \delta r$ are selected.
4) Among all selected combinations, the maximum PSNR value, $PSNRmax$, is determined.
5) A PSNR variation δq is set, and a minimum PSNR value, $PSNRmin = PSNRmax - \delta q$, is calculated.
6) Among all selected (Q_p, Q_{pText}) in step 3, those whose PSNR values are greater than $PSNRmin$ are chosen as candidates.
7) Select the candidate with the largest $d = Q_p - Q_{pText}$.

The image shown in Fig. 2(a) was compressed by AVC-C, AVC-I and JPEG-2000 with different parameters, and results are shown in Fig. 3.

The differences among the coders become more clear when text is visualized in details. Fig. 4 shows (a) AVC-C and (b) JPEG-2000 versions, for a zoomed region of a text letter border. A zoom of the pictorial region of the same image is shown in Fig. 4 (c) and (d).

4. CONCLUSIONS

The AVC-I is very effective for compound documents because of its intraframe prediction mode. With AVC-C, for the same bitrate, it is possible to improve significantly the quality of text regions, with little to negligible losses to the pictorial regions. The losses to pictorial regions yield an overall PSNR loss so small that AVC-C still outperforms JPEG-2000. Even though there is not an overall objective gain over AVC-I, the proposed AVC-C encodes text regions at higher quality and yields an overall higher quality to the image. Furthermore, the proposed AVC-C encoder is compatible with AVC-I decoder. Future works might consider extending the number of classes.

5. REFERENCES

[1] Final Draft International Standard of Joint Video Specification. ITU-T Recommendation H.264, ISO/IEC 14496-10 AVC. March 2003.

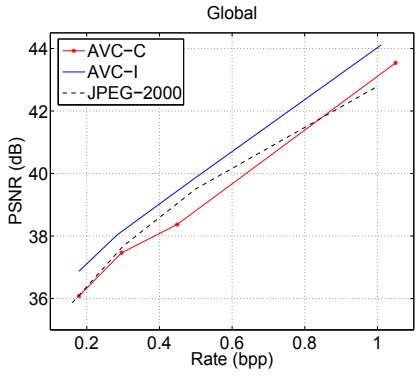
[2] I. E. G. Richardson, *H.264 and MPEG-4 Video Compression*, Wiley, 2003.

[3] T. Wiegand, G. J. Sullivan, G. Bjontegaard, and A. Luthra, "Overview of the H.264/AVC video coding standard", *IEEE Trans. on Circuits and Systes for Video Technology*, Vol. 13, No. 7, pp. 560-576, July 2003.

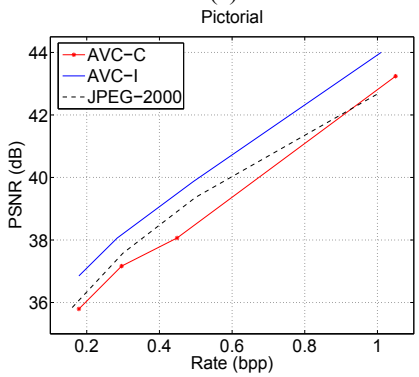
[4] G. Sullivan et al., "The H.264/AVC Advanced Video Coding Standard: Overview and Introduction to the Fidelity Range Extensions", *Proc. SPIE Conference on Applications of Digital Image Processing XXVII*, Special Session on Advances in the New Emerging Standard: H.264/AVC, August, 2004.

[5] J. Ostermann et. al.,"Video coding with H.264/AVC: tools, performance, and complexity", *IEEE Circuits and Systems Magazine*, pp. 7-28, Mar. 2004.

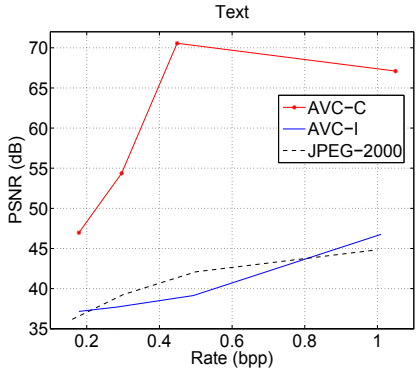
[6] D. Marpe, V. George, and T. Wiegand, "Performance comparison of intra-only H.264/AVC HP and JPEG2000 for a set of monochrome ISO/IEC test images", Contribution JVT ISO/IEC MPEG and ITU-T VCEG, JVT M-014, Oct. 2004.



(a)

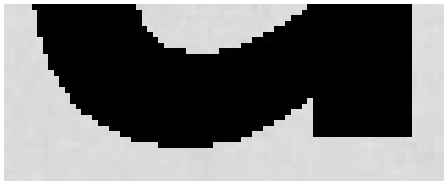


(b)

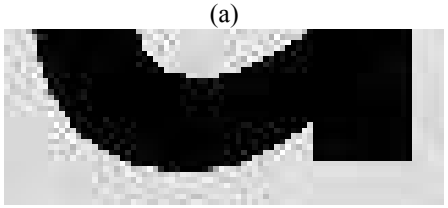


(c)

Fig. 3. Objective performance comparison between AVC-C, AVC-I and JPEG-2000 for "DSPG" image: (a) global PSNR; (b) pictorial regions PSNR; (c) text regions PSNR. Notice that text regions quality can be considerably improved with little global quality loss.



(a)



(b)

(c)

(d)

Fig. 4. Subjective performance comparison between AVC-C and JPEG-2000 for pictorial and text regions: (a) and (c) AVC-C encoded; (b) and (d) JPEG-2000 encoded. Bitrate is 0.16 bpp.

[7] D. Marpe, V. George, H. L. Cycon, and K. U. Barthel, "Performance evaluation of Motion-JPEG2000 in comparison with H.264/AVC operated in pure intra coding mode", Wavelet Applications in Industrial Processing, *Proceedings of the SPIE*, Vol. 5266, pp. 129-137, 2004.

[8] D. S. Taubman and M. W. Marcellin, JPEG2000: *Image Compression Fundamentals, Standards, and Practice*, Kluwer Academic, 2002.

[9] R. L. de Queiroz, R. S. Ortis, A. Zaghetto, and T. A. Fonseca, "Fringe benefits of the H.264/AVC", *Proc. of Intl. Telecom. Symp.*, Fortaleza, Brazil, pp. 208-212, Sep. 2006.

[10] R. de Queiroz, "Compression of compound documents", *Proc. ICIP*, Kobe, Japan, 25PS1.1, Oct. 1999.

[11] D. Marr, *Vision*, San Francisco, CA: Freeman, 1982.

[12] R. de Queiroz, Z. Fan, T. D. Tran, "Optimizing Block-Thresholding Segmentation for Multilayer Compression of Compound Images", *IEEE Trans. on Image Processing*, Vol. 9, No. 9, pp. 1461-1471, September 2000.

[13] K. Konstantinides, D. Tretter, "A JPEG Variable Quantization Method for Compound Documents", *IEEE Trans. on Image Processing*, Vol. 9, No. 7, pp. 1282-1287, July 2000.

[14] M. Ramos and R. de Queiroz, "Classified JPEG coding of mixed documents",*IEEE Trans. on Image Processing*, Vol. 9, pp. 716-720, April 2000.

[15] J. Fan, "Text extraction via an edge-bounded averaging and a parametric character model", *Proc. of the SPIE Document Recognition and Retrieval X*, Vol. 5010, pp. 8-19, Santa Clara, CA, USA, Jan. 2003.