# HEVC-BASED SCANNED DOCUMENT COMPRESSION

*Alexandre Zaghetto, Bruno Macchiavello and Ricardo L. de Queiroz*

Department of Computer Science
Universidade de Brasília, Brazil

## ABSTRACT

This paper proposes a hybrid pattern matching/transform-based compression engine for scanned compound documents. The novelty of this approach is demonstrated by using a modified version of the HEVC (High Efficiency Video Coding) Test Model as a compound document compressor, here conveniently referred to as HEDC (High Efficiency Document Coder). The proposed method uses segments of a document to create a video sequence, which is then encoded by HEDC. The idea is to explore interframe prediction as a pattern matching algorithm for coding units pre-classified as text; and intraframe prediction for coding units pre-classified as image. Results show that HEDC outperforms AVC-I, HEVC-I (H.264/AVC and HEVC operating in pure intra mode), H.264/AVC and JPEG2000 by up to 3.3, 2.5, 1.7 and 5 dB, respectively. Furthermore, for most documents the proposed method yields practically the same rate-distortion performance as regular HEVC, but is approximately 5% to 20% faster due to a pre-classification algorithm that prevents it of performing all possible inter/intra prediction tests for each prediction unit.

***Index Terms*—** Page processing, compound document compression, High Efficiency Video Coding, pattern matching.

## 1. INTRODUCTION

Scanned documents are either compressed as a continuous-tone picture, or they are binarized before compression. The binary document can then be compressed using any available two-level compression algorithm (such as JBIG [1] and JBIG2 [2]), or it may undergo character recognition [3]. Binarization may cause strong degradation to object contours and textures, such that, whenever possible, continuous-tone compression is preferred [4]. Examples of continuous-tone image compression algorithm are JPEG [5] or JPEG2000 [6, 7]. Some results point to the fact that the many coding advances brought into H.264/AVC operating in pure intra mode also made it a formidable compressor for still images [8, 9, 10]. Multilayer approaches such as the mixed raster content (MRC) imaging model [11, 12, 13, 14, 15] are also challenged by soft edges in scanned documents, often requiring pre- and post-processing [16].

Natural text along a document typically presents repetitive symbols such that dictionary-based compression methods become very efficient [17]. For continuous-tone imagery, the recurrence of similar patterns is illustrated in Fig. 1. Nevertheless, the development of an efficient dictionary-based encoder relying on continuous-tone pattern matching is a challenging problem. We propose an encoder that explores such a recurrence through the use of a page processing procedure, a coding unit pre-classification algorithm, pattern matching predictors and efficient transform encoding of the residual data.

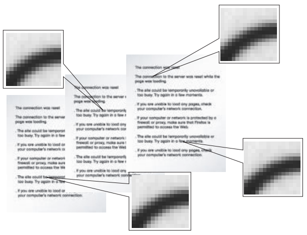


**Fig. 1**: Scanned documents usually present recurrent patterns.

## 2. THE PROPOSED METHOD

First, each scanned $H \times W$ pixels document is segmented into 16 ($H/4 \times W/4$ pixels) sub-pages. These sub-pages are organized as frames of a video sequence, which is further encoded through the proposed encoder. Figure 2 illustrates the page processing procedure.

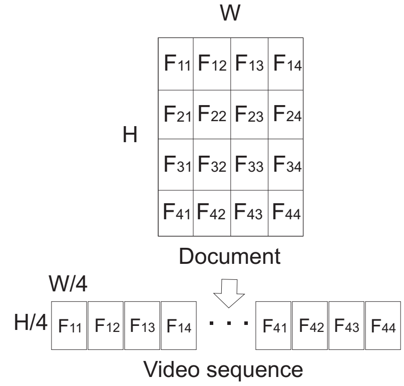
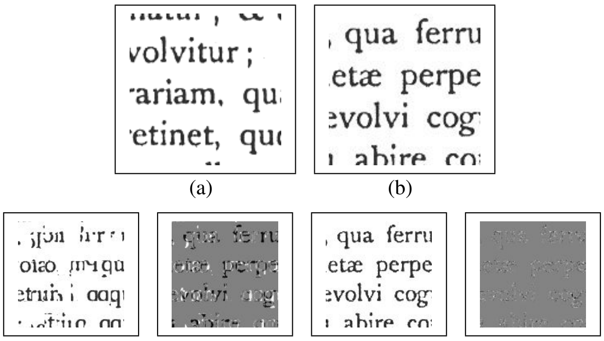


**Fig. 2**: Proposed page processing algorithm.

The encoder used here is based on pattern matching. Each frame is partitioned into coding units, which are further partitioned into prediction units. Each prediction unit in a frame is then matched to a pattern existing in a previous frame. Once a match is found, the matching pattern is used as a predictor and the prediction error (residue) is transform coded. Prediction units may be of different sizes and geometries. Figure 3 illustrates the effect of using the pattern matching algorithm with two different prediction units. Figures 3 (a) and (b) show examples of a reference and a current text area, respectively. Figures 3 (c) and (e) represent the predictions of the current text using $16 \times 16$ and $4 \times 4$ pixels prediction units, respectively. Figures 3 (d) and (f) are the corresponding residual data. Notice that the $4 \times 4$-pixel prediction generates a lower-energy residual, when compared to the $16 \times 16$, however, they require encoding

E-mail: {zaghetto, bruno, queiroz}@image.unb.br.

more reference vectors. This trade-off is addressed by an optimization algorithm.



(a) (b)

(c) (d) (e) (f)

**Fig. 3**: Illustration of pattern matching using interframe prediction: (a) reference text; (b) current text; (c) and (e) predicted text (block size: 16×16 and 4×4 pixels, respectively); and (d) and (f) prediction residue (block size: $16 \times 16$ and $4 \times 4$ pixels, respectively).

Since we are dealing with compound documents - images composed by a mixture of picture and text - we propose a coding unit pre-classification algorithm in order to avoid time consuming inefficient pattern matching. It is not likely that a prediction unit in a pictorial region of a frame would find any good match in a text region of a previously encoded frame. Pictorial regions would be better encoded through intraframe prediction instead.

The first step of the text/picture pre-classification algorithm binarizes the document using the Floyd-Steinberg dithering [18], described in the FSDITHERING procedure. $D(i, j)$ represent the image being binarized.

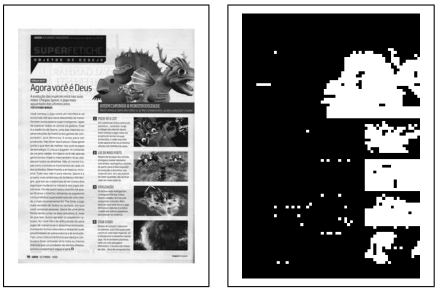
```
FSDITHERING
for $i = 0$ to $height$
    for $j = 0$ to $width$
        $p_{old} = D(i, j)$
        if $p_{old} \geq 127$
            $p_{new} = 255$
        else $p_{new} = 0$
        $D(i, j) = p_{new}$
        $\varepsilon = p_{old} - p_{new}$
        $D(i, j + 1) = \frac{7}{16}\varepsilon$
        $D(i + 1, j - 1) = \frac{3}{16}\varepsilon$
        $D(i + 1, j) = \frac{5}{16}\varepsilon$
        $D(i + 1, j + 1) = \frac{1}{16}\varepsilon$
```

Then, for each 64×64-pixel block a cross-correlation coefficient $C_r$ is calculated [19], according to Equation 1.

$$C_r = 1 - \frac{2}{64^2}\sum_{i=0}^{63}\sum_{j=0}^{63}[p(i, j) \text{ xor } p(i + 1, j)]. \quad (1)$$

If $Cr \geq 0.86$, the coding unit is classified as picture and is intraframe encoded; otherwise, the coding unit is a candidate for interframe prediction (pattern matching).

Figure 4 shows an example of compound document and the result of the described pre-classification algorithm. White pixels represent pictorial regions. To implement the proposed encoder, here conveniently called High Efficiency Document Coder (HEDC), we included the above described new features (page processing and coding unit pre-classification) in the recently proposed HEVC (High Efficiency Video Coding) [20] Test Model, which is meant to become the next generation video coder. This new standard is still under development by ISO/IEC and ITU-T. It was selected, because it incorporates a patter matching algorithm with variable block sizes. Each block is referred to as coding unit, which can be of sizes from $8 \times 8$ to $64 \times 64$ in a tree structure. Each coding unit can be further partitioned symmetrically or asymmetrically into prediction units. The predictions units can be coded using one of the 34 intra prediction modes, or using interframe prediction. HEVC also presents a quadtree structure transform coding with block size from $4 \times 4$ to $32 \times 32$. The best block partitions and coding modes are determined in a rate-distortion sense.



(a) (b)

**Fig. 4**: Result of pre-classification algorithm. White pixels represent pictorial regions and are intraframe encoded. Black regions are candidate for interframe prediction (pattern matching).

## 3. RESULTS

In our tests, the documents shown in Fig. 5 are compressed using JPEG2000 (JP2), AVC-I and HEVC-I (H.264/AVC and HEVC operating in pure intraframe mode, respectively), conventional H.264/AVC [21] (AVC), conventional HEVC, and the proposed HEDC. For JPEG2000, AVC-I and HEVC-I compression, the frames are separately encoded as still images. As for H.264/AVC, HEVC and HEDC the first frame of the sequence is encoded as an I-frame (only intraframe prediction modes are used) and all the remaining frames are encoded as P-frames (past frames are used as reference by the interframe prediction). The number of reference frames, $R_f$, and the search range, $S_r$, where set to 4 and 64, respectively. Figure 6 shows PSNR plots comparing the above mentioned encoders. The PSNR was calculated using the global mean square error (MSE), instead of an average of the PSNR for each frame. Unlike conventional video sequences, here the frames can have very different characteristics, *e. g.*, one frame can be completely pictorial and the next frame completely composed of text regions, therefore an overall PSNR is a better quality indicator than average PSNR.

All encoders that use interframe prediction outperform those that do not; namely, AVC-I, HEVC-I and JPEG2000. This corroborate previous studies that point interframe prediction as an efficient pattern matching algorithm for text compression in scanned documents [21]. For some images, HEVC-I may present a similar performance compared to regular H.264/AVC. This indicates the effectiveness of the improvements brought into this under development standard. However, regular HEVC outperforms all other codecs for all tested documents. Compared to HEVC, the proposed HEDC presents almost no PSNR difference for four of the tested documents, as can be seen in Table 1. The most noticeble drops in

**Table 1**: Average performance comparison between HEVC and HEDC. PSNR loss and complexity reduction of HEDC are 0.10 dB and 12.94%, respectvely.

| Document | ΔPSNR (dB) | Complexity reduction (%) |
|---|---|---|
| carta | 0 | 8.36 |
| IEEE1 | -0.25 | 14.36 |
| IEEE2 | -0.21 | 13.93 |
| scientific | -0.06 | 14.06 |
| spectrum | -0.04 | 5.24 |
| spore | -0.02 | 21.67 |
| Average: | **-0.10** | **12,94** |

PSNR may be observed in documents "IEEE1" and "IEEE2", where the average PSNR differences are 0.25 and 0.21, respectively. These differences, however, do not affect subjective quality. Furthermore, HEDC significantly reduces the encoding time compared to regular HEVC. The average time reduction is 12.94%, while the upper and lower bounds are 21.67% and 5.24%, respectively.

Figure 7 shows a subjective comparison between the proposed method and JPEG2000. For both regions, pictorial and text, the proposed encoder presents superior quality. It preserves more adequately the edges for both text and pictorial data, with less ringing effect around text regions.

## 4. CONCLUSION

In this paper we proposed a scanned compound document encoder that uses: (a) a page processing procedure; (b) a text/picture pre-classification algorithm; (c) pattern matching with variable block sizes for text regions; (c) intra coding for pictorial regions; and (d) efficient transform encoding of residual data. This encoder was implemented using the HEVC Test Model as framework. Results show that HEDC consistently outperforms AVC-I, HEVC-I, H.264/AVC and JPEG2000. Furthermore, for most documents the proposed method presents practically the same rate-distortion performance as regular HEVC, but a significant complexity reduction is achieved due to the proposed pre-classification algorithm.

## 5. REFERENCES

[1] JBIG, "Information Technology - Coded Representation of Picture and Audio Information - Progressive Bi-level Image Compression. ITU-T Recommendation T.82," March 1993.

[2] JBIG2, "Information Technology - Coded Representation of Picture and Audio Information - Lossy/Lossless Coding of Bi-level Images. ITU-T Recommendation T.88," March 2000.

[3] S. Mori, C.Y. Suen, and K.; Yamamoto, "Historical review of OCR research and development," *Proceedings of the IEEE*, vol. 80, no. 7, pp. 1029–1058, Jul. 1992.

[4] R. L. de Queiroz, *Compressing Compound Documents, in The Document and Image Compression Handbook*, by M. Barni, Marcel-Dekker, EUA, 2005.

[5] W. B. Pennebaker and J. L. Mitchell, *JPEG Still Image Data Compression Standard*, Chapman and Hall, 1993.

[6] JPEG, "Information Technology - JPEG2000 Image Coding System - Part 1: Core Coding System. ISO/IEC 15444-1," 2000.

[7] D. S. Taubman and M. W. Marcellin, *JPEG 2000: Imagem Compression Fundamentals, Standards and Practice*, Kluwer Academic, EUA, 2002.

[8] D. Marpe, V. George, and T. Wiegand, "Performance Comparison of Intra-only H.264/AVC and JPEG2000 for a Set of Monochrome ISO/IEC Test Images," *Contribution JVT ISO/IEC MPEG and ITU-T VCEG, Doc. JVT M-014*, Oct. 2004.

[9] R. L. de Queiroz, R. S. Ortis, A. Zaghetto, and T. A. Fonseca, "Fringe Benefits of the H.264/AVC," *Proceedings of International Telecommunications Symposium*, pp. 208–212, Sept. 2006.

[10] A. Zaghetto and R. de Queiroz, "Segmentation-driven Compound Document Coding based on H.264/AVC-INTRA," *IEEE Transactions on Image Processing*, vol. 16, no. 7, pp. 1755–1760, Julho 2007.

[11] MRC, "Mixed Raster Content (MRC). ITU-T Recommendation T.44.," 1999.

[12] R. L. de Queiroz, R. Buckley, and M. Xu, "Mixed Raster Content (MRC) Model for Compound Image Compression," *Proceedings of SPIE Visual Communications and Image Processing*, vol. 3653, pp. 1106–1117, Jan. 1999.

[13] Patrick Haffner, Paul G. Howard, Patrice Simard, Yoshua Bengio, and Yann Lecun, "High Quality Document Image Compression with DjVu," *Journal of Electronic Imaging*, vol. 7, pp. 410–425, 1998.

[14] G. Feng and C. A. Bouman, "High-quality MRC Document Coding," *IEEE Transactions on Image Processing*, vol. 15, no. 10, pp. 3152–3169, Oct. 2006.

[15] A. Zaghetto, R. L de Queiroz, and D. Mukherjee, "MRC Compression of Compound Documents using Threshold Segmentation, Iterative Data-filling and H.264/AVC-INTRA," *Proceedings of Indian Conference on Computer Vision, Graphics and Image Processing*, Dezembro 2008.

[16] A. Zaghetto and R. L de Queiroz, "Pre- and postprocessing for multilayer compression of scanned documents," *Journal of Electronic Imaging*, , no. 20, pp. 043005, Oct. 2011.

[17] N. Francisco, N. Rodrigues, E. da Silva, M. de Carvalho, S. de Faria, and V. da Silva, "Scanned compound document encoding using multiscale recurrent patterns," *IEEE Transactions on Image Processing*, vol. 19, no. 10, pp. 2712–2724, Apr. 2010.

[18] Robert W. Floyd and Louis Steinberg, "An Adaptive Algorithm for Spatial Greyscale," *Proceedings of the Society for Information Display*, vol. 17, no. 2, pp. 75–77, 1976.

[19] Qing Wang, Zheru Chi, and Rongchun Zhao, "Hierarchical content classification and script determination for automatic document image processing," in *Proceedings of the 16 th International Conference on Pattern Recognition*, Washington, DC, USA, 2002, pp. 30077–.

[20] K. Ugur, K. Andersson, A. Fuldseth, G. Bjontegaard, L.P. Endresen, J. Lainema, A. Hallapuro, J. Ridge, D. Rusanovskyy, Cixun Zhang, A. Norkin, C. Priddle, T. Rusert, J. Samuelsson, R. Sjoberg, and Zhuangfei Wu, "High Performance, Low Complexity Video Coding and the Emerging HEVC Standard," *IEEE Transactions on Circiuts and Systems for Video Technology*, vol. 20, no. 12, pp. 1688–1697, Dec. 2010.

[21] A. Zaghetto and R.L. de Queiroz, "High quality scanned book compression using pattern matching," in *Image Processing (ICIP), 2010 17th IEEE International Conference on*, Sept. 2010, pp. 2165–2168.

(a) (b) (c) (d) (e) (f)

**Fig. 5**: Test set used in our experiments: (a) “carta” (size: 2304 × 1792 pixels); (b) “IEEE1” (size: 3328 × 2304 pixels); (c) “IEEE2” (size: 3328 × 2304 pixels); (d) “scientific” (size: 2304 × 1792 pixels); (e) “spectrum” (size: 2048 × 1792 pixels); and (f) “spore” (size: 1536 × 1024 pixels).

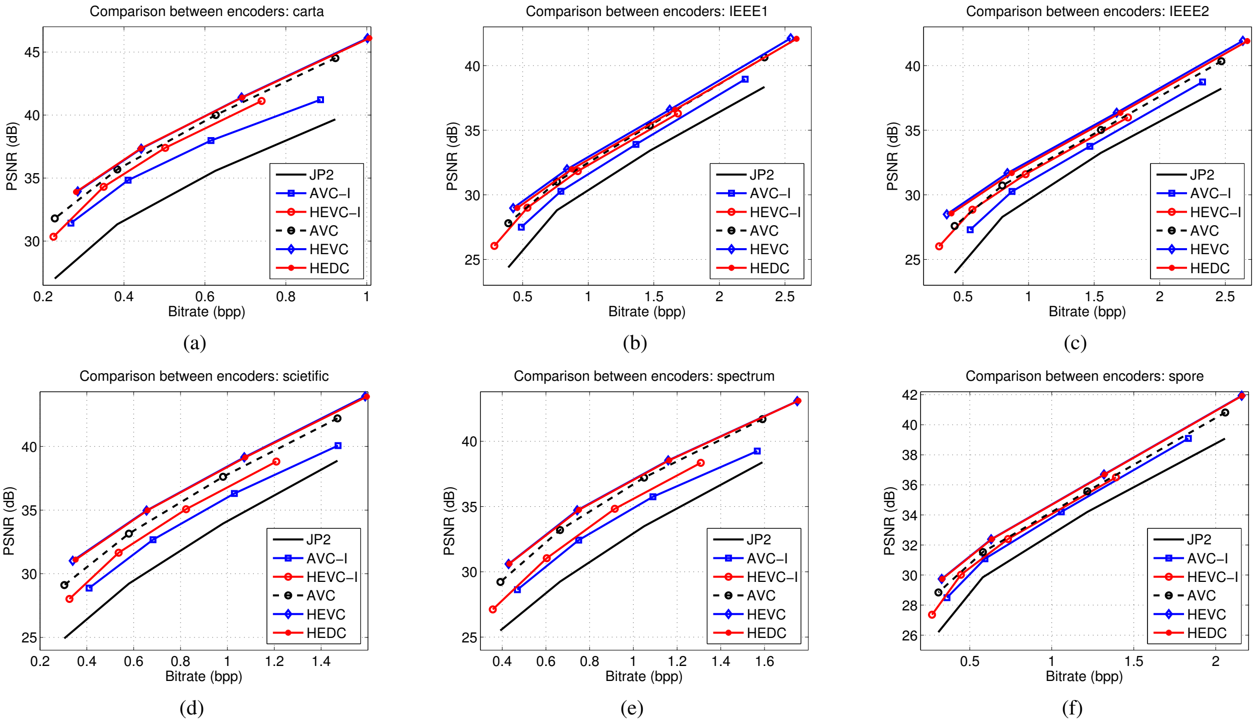
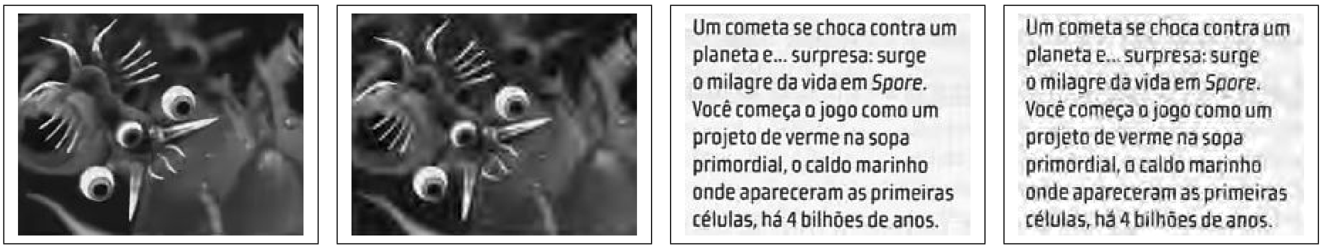


**Fig. 6**: PSNR plots for the test set shown in Fig. 5: (a) “carta”; (b) “IEEE1”; (c) “IEEE2”; (d) “scientific”; (e) “spectrum”; and (f) “spore”.



Um cometa se choca contra um
planeta e... surpresa: surge
o milagre da vida em *Spore*.
Você começa o jogo como um
projeto de verme na sopa
primordial, o caldo marinho
onde apareceram as primeiras
células, há 4 bilhões de anos.

Um cometa se choca contra um
planeta e... surpresa: surge
o milagre da vida em *Spore*.
Você começa o jogo como um
projeto de verme na sopa
primordial, o caldo marinho
onde apareceram as primeiras
células, há 4 bilhões de anos.

(a) (b) (c) (d)

**Fig. 7**: Subjective comparison between the proposed HEDC and the state-of-the-art continuous-tone still image compressor JPEG2000. These two encoders presented the best and the worst observed rate-distortion performance, respectively. The images represent zoomed parts of “spore” document compressed at approximately 0.3 bits/pixels: (a) HEDC-encoded pictorial region; (b) JPEG2000-encoded pictorial region; (c) HEDC-encoded text region; and (d) JPEG2000-encoded text region.