A Theoretical Framework for Data-Hiding in Digital and Printed Text Documents

R. Villán, S. Voloshynovskiy, F. Deguillaume, Y. Rytsar,
O. Koval, E. Topak, E. Rivera, and T. Pun

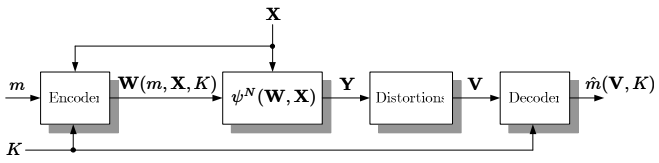
Stochastic Image Processing Group, Computer Vision and Multimedia Laboratory, University of Geneva

Introduction

- Is the problem of text data-hiding too difficult to solve?
- Answer depends on the application requirements:
  - If *robust data-hiding* is required (e.g. copyright protection): probably YES, the attacker can always use Optical Character Recognition.
  - If either *semi-fragile* or *fragile data-hiding* is required (e.g. identification, authentication and tamper proofing): NO
- Goals:
  - New theoretical framework for the text data-hiding problem.
  - New semi-fragile text data-hiding method, *color quantization*, that is fully automatable, has high information embedding rate, is resistant to printing and scanning, and can be applied to both digital and printed text documents.

1

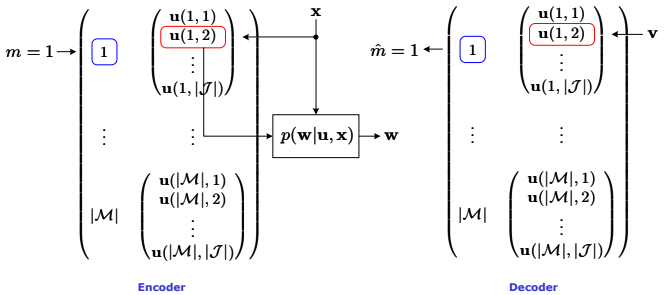
Text Data-Hiding as a Gel'fand-Pinsker (G-P) Problem [1]



- The text is represented by $\mathbf{X}$. Each component X_n, $n = 1, 2, \ldots, N$ of $\mathbf{X}$ represents *one character* from this text.
- A character is as an element from a given language alphabet (e.g. {A, B, …,Z}). More precisely, a character X_n is a *data structure consisting of multiple component fields* (features): name, shape, position, orientation, size, color, etc.

2

Gel'fand-Pinsker Encoder and Decoder



3

Example of G-P text data-hiding

- Consider the so-called Scalar Costa Scheme (SCS) [2]. The auxiliary random variable U is approximated by:

$$U = W + \alpha' X = \alpha' Q_m(X)$$

- The stego text Y is obtained as:

$$Y = W + X = \alpha' Q_m(X) + (1 - \alpha') X \qquad (*)$$

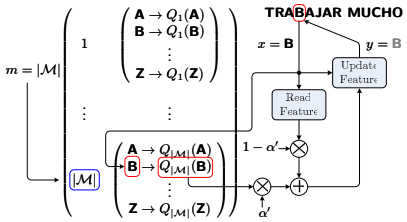
compensation parameter factor

high rate scalar quantizer

- For a practical implementation based on the SCS, just select a character feature (e.g. color) and use it as the cover character X.

4

Example of G-P text data-hiding (cont.)



- Generalization: $Q_m(\cdot) \rightarrow \mathbf{Q}_m(\cdot)$ E.g.: $\mathbf{Q}_m(\mathbf{TRABAJAR})$
- *Open space methods* and *character feature methods* are particular cases of the vector and the scalar schemes, respectively.

5

Practical Implementation of G-P Text Data-Hiding

- Color quantization: the stego text is obtained via (*), where the character feature to quantize is *color*:

VAMOS A TRABAJAR → **VAMOS A TRABAJAR**

- Main idea: quantize the color of each character in such a manner that the HVS cannot make the difference between original and quantized characters, but it is possible for an specialized reader.
- Embedding rate: 1-2 bits per character.
- Automation: correct character segmentation is needed for decoding; however OCR is not necessary.
- *Two-level* or *multilevel* quantizers can be used.

6

Practical Implementation of G-P Text Data-Hiding

- Halftone quantization: it exploits the fact that there exists a number of choices for the halftone screen leading to the same gray shade.
- Typical halftone screen characteristics that can be exploited are: *screen angle* and *screen dot shape* (elliptical, round, square).

$\mathbf{Q}_0(\mathbf{T})$ $\mathbf{Q}_1(\mathbf{T})$

7

Conclusions and Future Work

- New theoretical framework for data-hiding in digital and printed text documents: *G-P text data-hiding*.
- Main idea: consider a text character as a *data structure consisting of multiple features*.
- We presented *color quantization* as a new method for data-hiding in digital and printed text documents.
- Future work: Experimental results on the robustness of color quantization against printing and scanning.
- References:

1. Gel'fand, S., Pinsker, M.: Coding for channel with random parameters. Problems of Control and Information Theory 9 (1980) 19–31
2. Eggers, J., Su, J., Girod, B.: A blind watermarking scheme based on structured codebooks. In: Secure Images and Image Authentication, IEE Colloquium, London, UK (2000) 4/1–4/6