Text Data-Hiding for Digital and Printed Documents: Theoretical and Practical Considerations

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

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

Agenda

- **Introduction**
- **Text Data-Hiding as a Gel'fand-Pinsker (G-P) Problem**
  - § Encoder and Decoder
- **Practical Implementation of G-P Text Data-Hiding**
  - § Color Quantization
  - § Halftone Quantization
  - § Error Control Coding for Print-and-Scan Channels
- **Experimental Results**
- **Conclusions**

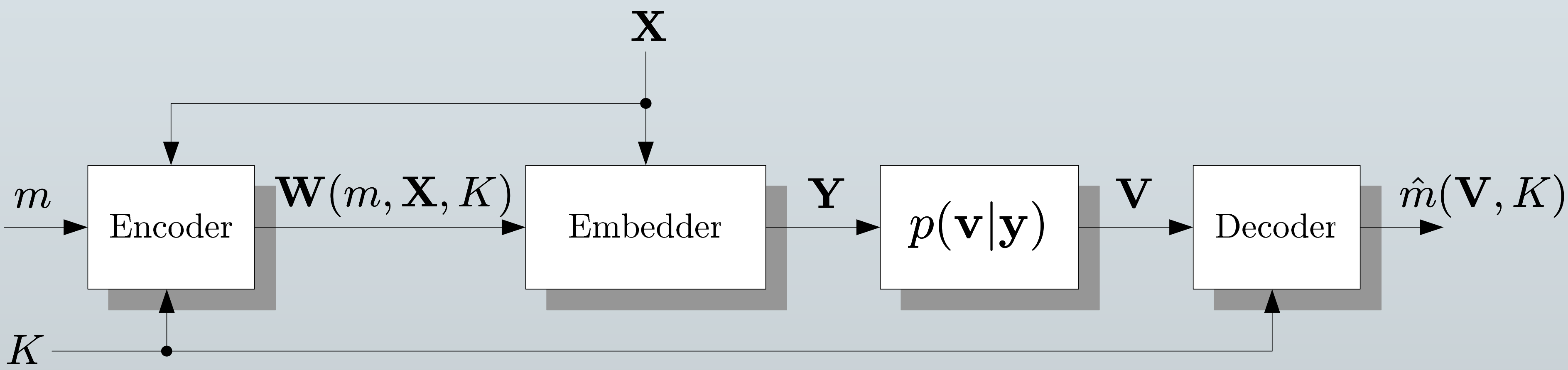
Introduction

- Text documents are omnipresent everyday: newspapers, books, web pages, contracts, advertisements, checks, identification documents, etc.
- What about the text data-hiding problem?
- Four major groups of methods for data-hiding in text documents: *syntactic methods*, *semantic methods*, *open space methods*, *character feature methods*.
- Some drawbacks: not suitable for all type of text documents (contracts, identity documents, literary texts), need human supervision, low information embedding rates, not robust against printing and scanning.

Introduction

- Text data-hiding, what for?
- Difficult for *robust data-hiding* applications (e.g. copyright protection) since the attacker can always use Optical Character Recognition (OCR).
- Possible for *semi-fragile* or *fragile data-hiding* applications (e.g. identification, authentication, tamper proofing, copy protection).
- 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 simultaneously to both digital and printed text documents.

Text Data-Hiding as a Gel'fand-Pinsker Problem

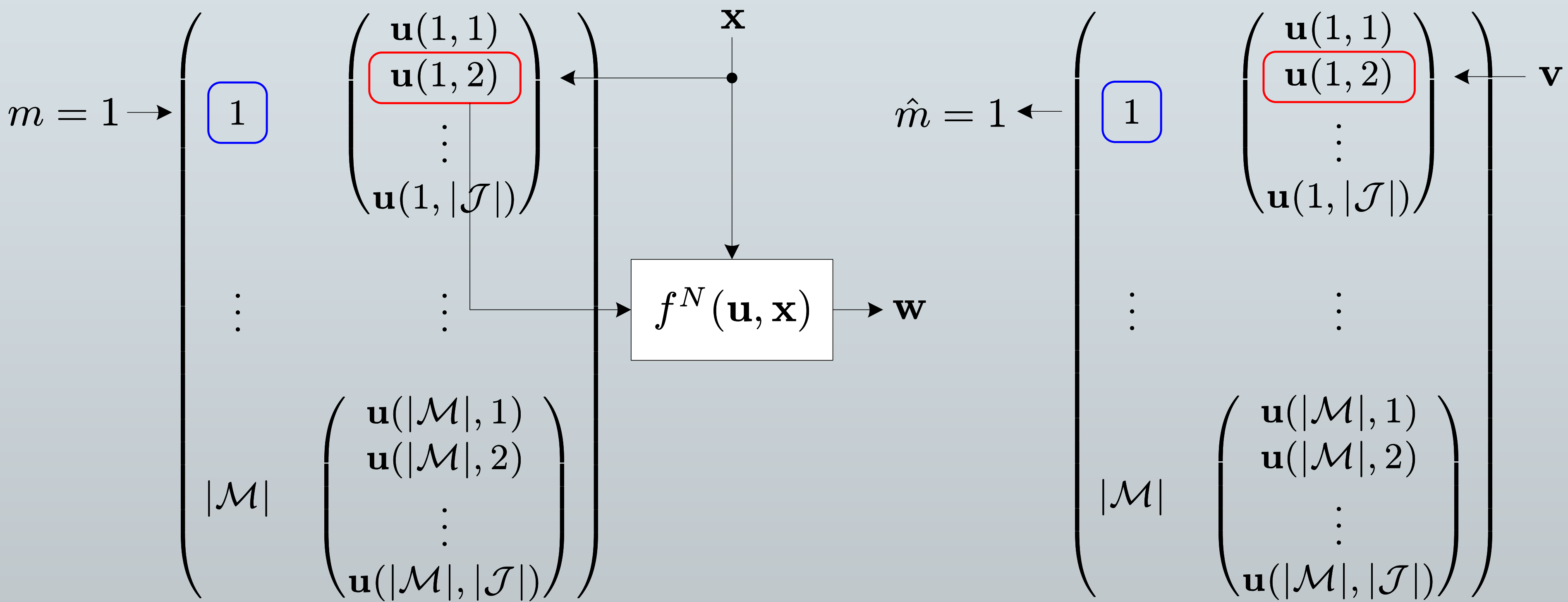


$$m \in \mathcal{M} = \{1,2,\ldots,|\mathcal{M}|\},\ |\mathcal{M}| = 2^{NR},\ K \in \mathcal{K} = \{1,2,\ldots,|\mathcal{K}|\}$$

$$\mathbf{X} = (X_1,\ldots,X_N) \in \mathcal{X}^N,\ \mathbf{W} \in \mathcal{W}^N,\ \mathbf{Y} \in \mathcal{Y}^N,\ \mathbf{V} \in \mathcal{V}^N$$

- A character X_n is a data structure consisting of multiple quantifiable component fields (features): shape (geometric definition), position, orientation, size, color, etc.

Encoder and Decoder



Encoder

$$(\mathbf{u}(m,j),\mathbf{x}) \in A_\epsilon^{*(N)}(U,X)$$

Decoder

$$(\mathbf{u}(\hat{m},j),\mathbf{v}) \in A_\epsilon^{*(N)}(U,V)$$

Text Data-Hiding as a Gel'fand-Pinsker Problem

- Costa considered the G-P problem for Gaussian variables.
- Costa's result still makes use of random codebooks with an exponential number of codewords in order to achieve capacity.
- To reduce the complexity of practical implementations, the use of structured codebooks has been proposed.
- For example, in the so-called Scalar Costa Scheme (SCS) the auxiliary random variable U is approximated by:

compensation parameter factor

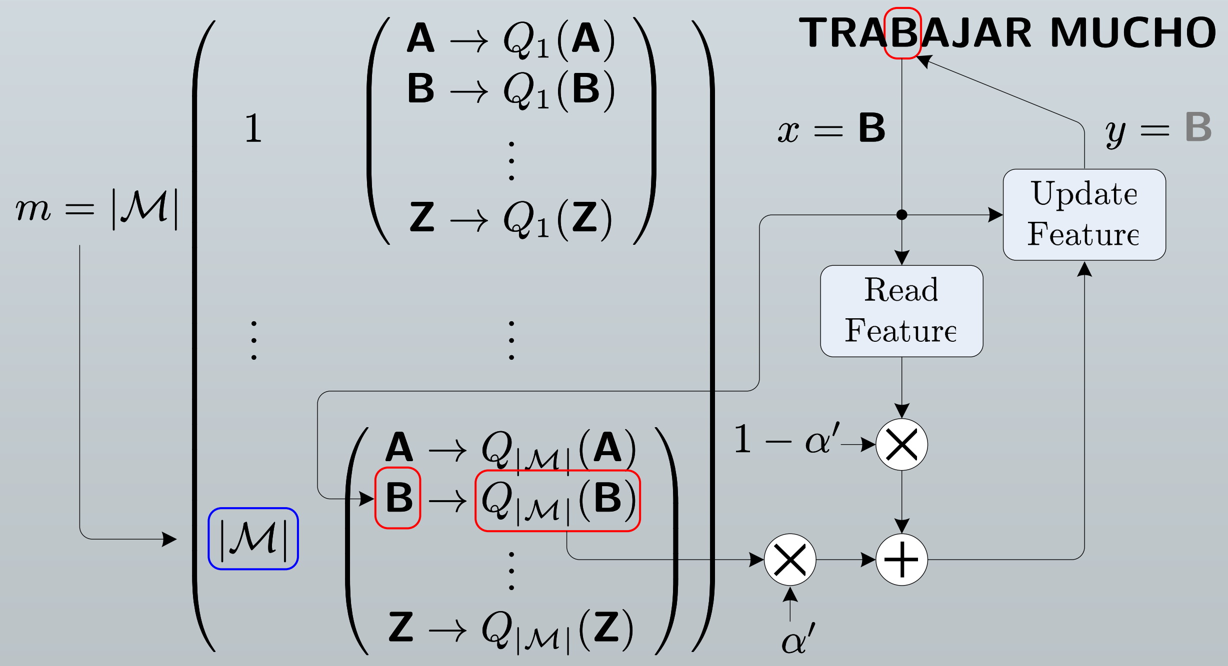
$$U = W + \alpha' X = \alpha' Q_m(X)$$

high rate scalar quantizer

- The resulting stego data is:

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

- Example: just select a character feature (e.g. color) and use it as the cover character X in («).



Text Data-Hiding as a Gel'fand-Pinsker Problem

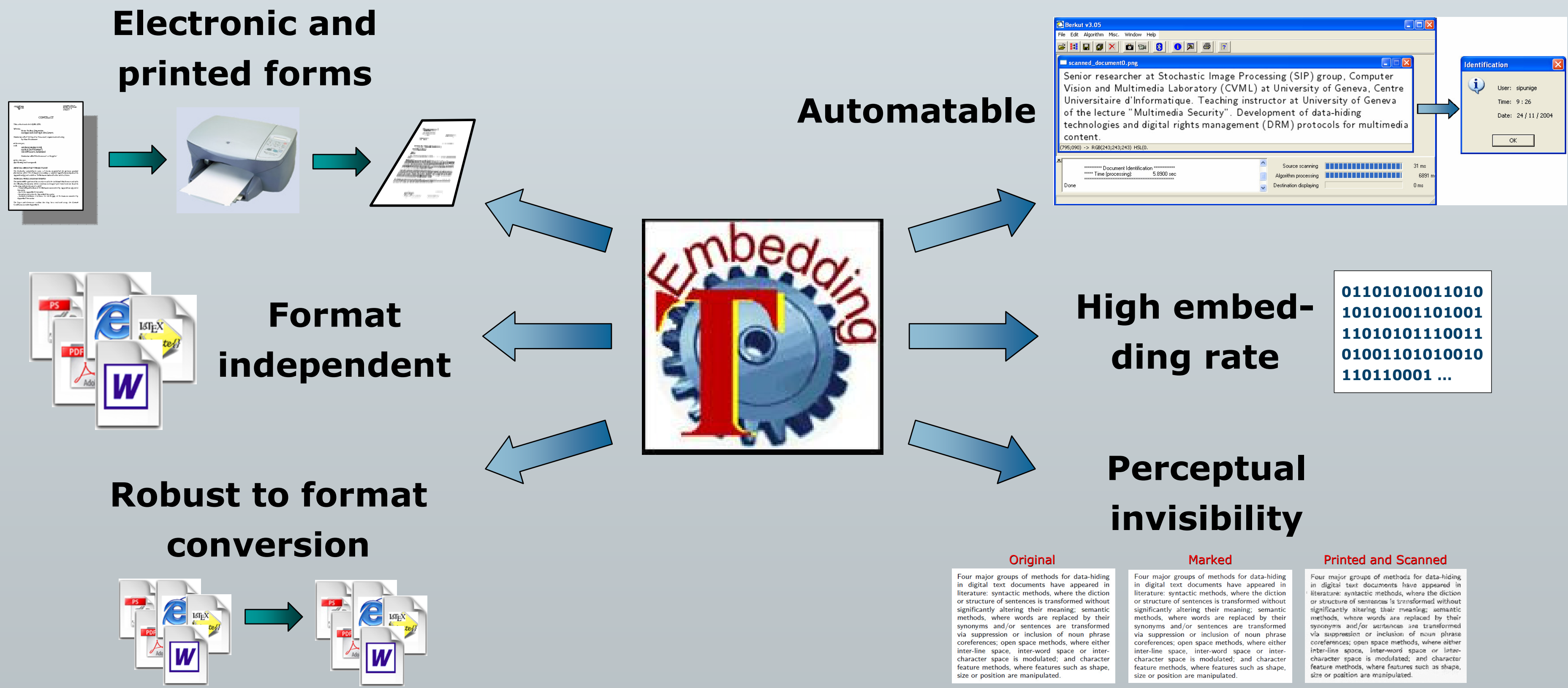
- Generalization 1. Select simultaneously *more than one character feature*, e.g. size and color. The quantizer $Q_m(\cdot)$ in («) becomes a vector quantizer $\mathbf{Q}_m(\cdot)$ acting on the selected character features. Main advantage: higher data embedding rate.
- Generalization 2. *Vector case*:

$$\mathbf{Y} = \mathbf{W} + \mathbf{X} = \alpha' \mathbf{Q}_m(\mathbf{X}) + (1 - \alpha')\mathbf{X} \qquad (« «)$$

  E.g. ($N = 8$) a codebook would contain the entry $\mathbf{Q}_m($**TRABAJAR**$)$ corresponding to message m and the group of characters $\mathbf{x} =$ **TRABAJAR**.
- *Open space methods* (feature *position*) and *character feature methods* are particular cases of (« «) and («), respectively.

Practical Implementation of G-P Text Data-Hiding

- Main requirements for a semi-fragile text data-hiding method:



Color Quantization

- The stego text is obtained via («), where $\alpha' = 1$ and the character feature to quantize is color:

VAMOS A TRABAJAR

$0\ 1\ 0\ 1\ 1\ \ 0\ \ 0\ 1\ 0\ 0\ 0\ 1\ 0\ 1$

VAMOS A TRABAJAR

- Main idea: quantize the color intensity of each character in such a way 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.

Example: Two-Level Color Quantization

Original

Four major groups of methods for data-hiding in digital text documents have appeared in literature: syntactic methods, where the diction or structure of sentences is transformed without significantly altering their meaning; semantic methods, where words are replaced by their synonyms and/or sentences are transformed via suppression or inclusion of noun phrase coreferences; open space methods, where either inter-line space, inter-word space or inter-character space is modulated; and character feature methods, where features such as shape, size or position are manipulated.

Marked

Four major groups of methods for data-hiding in digital text documents have appeared in literature: syntactic methods, where the diction or structure of sentences is transformed without significantly altering their meaning; semantic methods, where words are replaced by their synonyms and/or sentences are transformed via suppression or inclusion of noun phrase coreferences; open space methods, where either inter-line space, inter-word space or inter-character space is modulated; and character feature methods, where features such as shape, size or position are manipulated.

Printed and Scanned

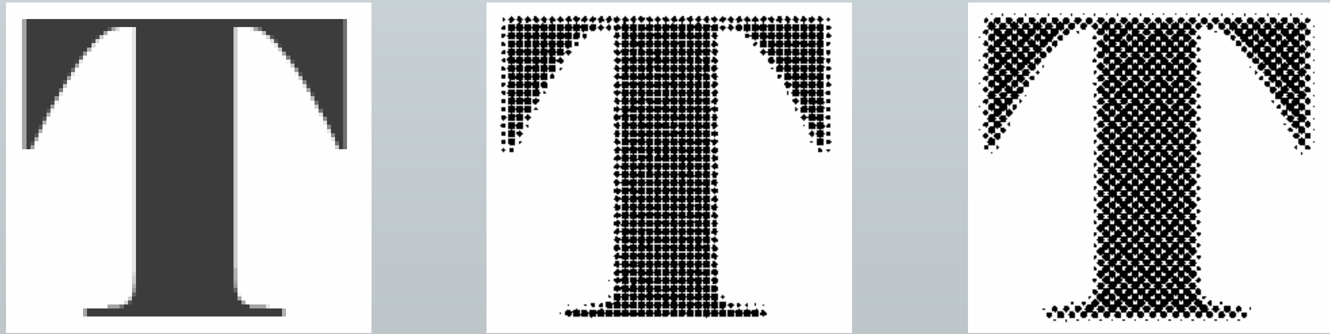
Four major groups of methods for data-hiding in digital text documents have appeared in literature: syntactic methods, where the diction or structure of sentences is transformed without significantly altering their meaning; semantic methods, where words are replaced by their synonyms and/or sentences are transformed via suppression or inclusion of noun phrase coreferences; open space methods, where either inter-line space, inter-word space or inter-character space is modulated; and character feature methods, where features such as shape, size or position are manipulated.

483 characters

Payload = 20 bytes

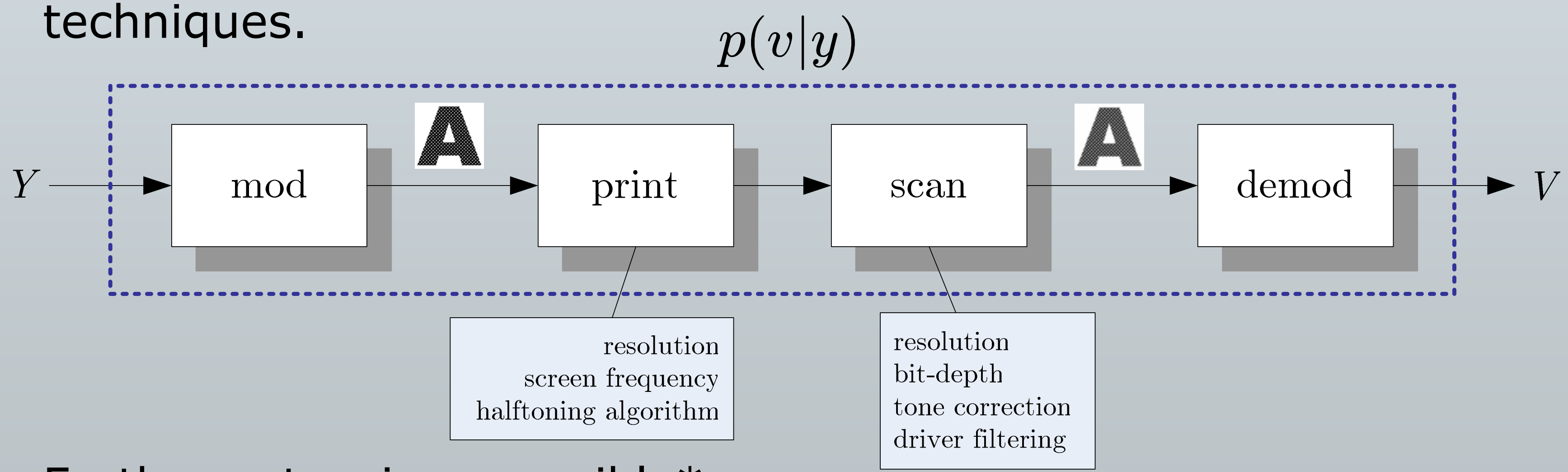
R=1/3 bits per character

- **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).



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

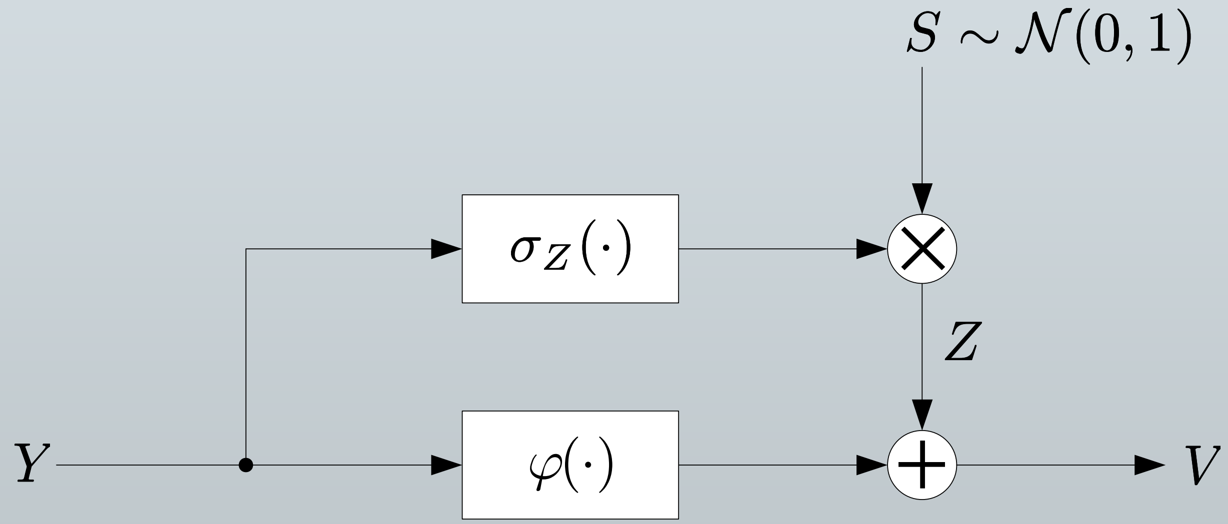
- An outer layer of coding can be used taking into account the channel formed by the *quantization encoder*, the *print-and-scan channel*, and the *quantization decoder*.
- Some modifications to get full benefit of soft-decision decoding techniques.



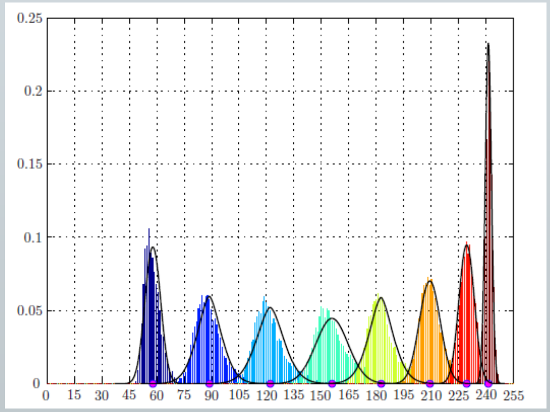
- Further extensions possible*.

* Pedro Comesaña, Fernando Pérez-González, and Frans M. J. Willems, "Applying Erez and Ten Brink's Dirty Paper Codes to Data-Hiding," in Proceedings of SPIE-IS&T Electronic Imaging 2005, 5681, pp. 298-307, San Jose, California, USA, January 2005.

- Effective coding techniques for print-and-scan channels have already been studied in the context of 2D bar codes*.



GGD approximation



- If a multilevel quantizer is used then a **multilevel encoder** together with a **multistage decoder** can be designed for the overall channel.

* R. Villán, S. Voloshynovskiy, O. Koval, and T. Pun, “Multilevel 2D Bar Codes: Towards High Capacity Storage Modules for Multimedia Security and Management,” in Proceedings of SPIE-IS&T Electronic Imaging 2005, 5681, pp. 453–464, (San Jose, USA), January 16–20 2005.

Experimental Results

- We implemented a *two-level color quantization scheme* (for electronic and printed documents):

  Printing and scanning at 600 dpi.

  Extraction process: segmentation of characters, demodulation of character features (color), and quantization-based decoding.

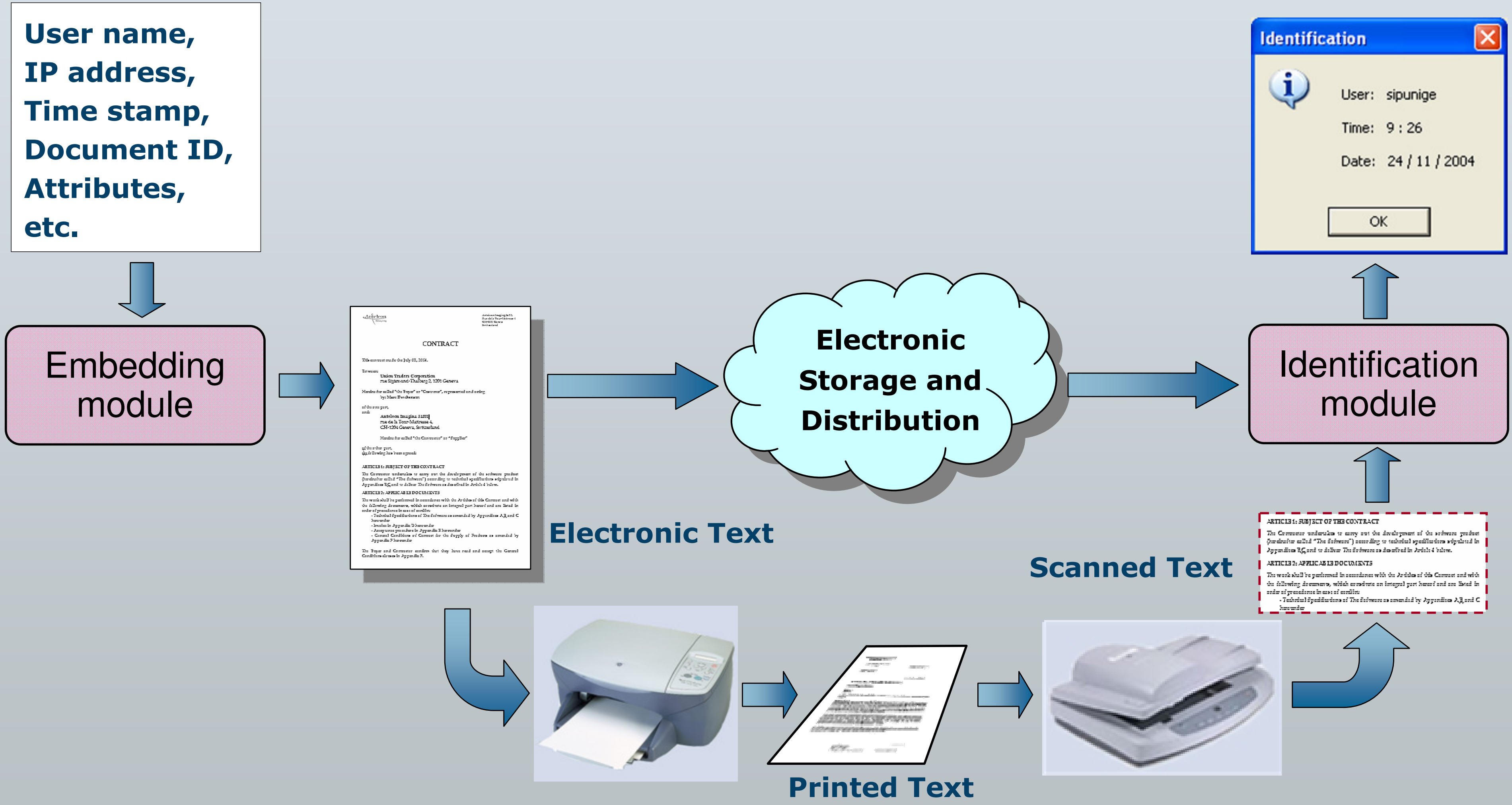
  Two choices for demodulation: computation of the average luminance, analysis of halftone pattern (better results).

$Q_0(x)$	$Q_1(x)$	**Average prob. of error**
0	41	0.327
0	46	0.201
0	51	0.077
0	56	0.029
0	61	0.015
0	66	0.006

text length = 4104 characters

$0 \rightarrow$ black

Example of Application: Document Identification



Conclusions

- 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 quantifiable features.
- Open space methods and character feature methods are particular cases of a general quantization-based text data-hiding technique.
- We presented *color quantization* as a new method for data-hiding in digital and printed text documents.
  - The experimental work confirmed this method has high perceptual invisibility, high information embedding rate, and is fully automatable.
  - Suitable for document identification, authentication, tamper proofing applications, and copy detection.