Tamper-proofing of Electronic and Printed Text Documents via Robust Hashing and Data-Hiding

R. Villán, S. Voloshynovskiy, O. Koval, F. Deguillaume, and T. Pun

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

Outline

- **Introduction**
- **Self-Authentication of Documents**
- **Gel'fand-Pinsker Text Data-Hiding**
- **Robust Hashing of Text Documents**
  - OCR + MAC Text Hashing
  - Random Tiling Text Hashing
- **Experimental Results**
- **Conclusions**

Introduction

- Problems (for both electronic and printed docs):
  - **Text document authentication:**
    Is the document authentic? **(global decision)**
  - **Text document tamper-proofing:**
    Is the document locally modified? **(local decision)**

- Possible solution:
  Generation of the document's hash based on a secret key K_H**.**

Introduction

Open issues:

- **Hash storage:**

  in a remote electronic database,

  onto the document itself using special means such as 2D bar codes, special inks or crystals, magnetic stripes, memory chips, etc.,

  onto the document's content itself using data-hiding techniques (self-authentication).

- **Hash requirements:**

  robustness to legitimate modifications,

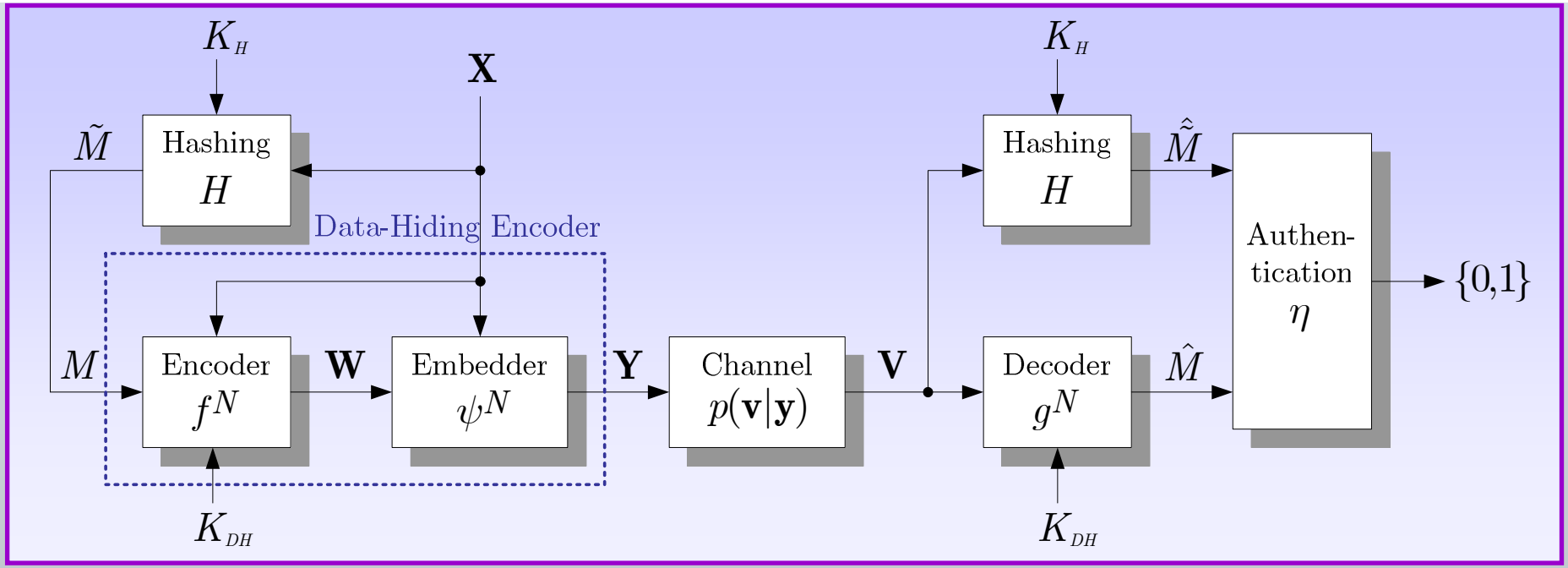
  sensitivity to attacks,

  security.

Introduction

- Main advantages of the self-authentication approach:
  - authentication of the document is performed directly without accessing a hash database,
  - hash cannot be easily separated from the document like it is, if a dense 2D bar code is used for storing the hash.
- Main concerns of the self-authentication approach:
  - limited data storage rate offered by current text data-hiding methods,
  - lack of reliable and secure robust text hashing functions.
- Main goal of our study:
  - to address the problem of limited data storage capacity of current text data-hiding technologies,
  - to study the properties of possibly good candidates for robust text hashing.

Self-Authentication of Documents



$$K_{DH} \in \mathcal{K}_{DH} = \{1, 2, \ldots, |\mathcal{K}_{DH}|\} \quad M \in \mathcal{M} = \{1, 2, \ldots, |\mathcal{M}|\},\ |\mathcal{M}| = 2^{NR_{DH}}$$

$$K_H \in \mathcal{K}_H = \{1, 2, \ldots, |\mathcal{K}_H|\} \quad \tilde{M} \in \tilde{\mathcal{M}} = \{1, 2, \ldots, |\tilde{\mathcal{M}}|\},\ |\tilde{\mathcal{M}}| = 2^{NR_H}$$

$$\mathbf{X} \in \mathcal{X}^N, \mathbf{W} \in \mathcal{W}^N, \mathbf{Y} \in \mathcal{Y}^N, \mathbf{V} \in \mathcal{V}^N \qquad p(\mathbf{v}|\mathbf{y}) = \prod_{i=1}^{N} p_{V|Y}(v_i|y_i)$$

Self-Authentication of Documents

- A self-authentication system consists of 3 parts:
  - hash function H with rate R_H,
  - text data-hiding encoder (f^N, ψ^N) and decoder g^N with rate R_{DH},
  - authentication function η .
- Hashing:

$$H : \mathcal{K}_H \times \mathcal{X}^N \rightarrow \tilde{\mathcal{M}}$$

- Text data-hiding:
  - Encoder:

$$f^N : \mathcal{M} \times \mathcal{X}^N \times \mathcal{K}_{DH} \rightarrow \mathcal{W}^N$$

  - Embedder:

$$\psi^N : \mathcal{W}^N \times \mathcal{X}^N \rightarrow \mathcal{Y}^N$$

  - Decoder:

$$g^N : \mathcal{V}^N \times \mathcal{K}_{DH} \rightarrow \mathcal{M}$$

Distortion metric:

$$d^N(\mathbf{x}, \mathbf{y}) = \frac{1}{N} \sum_{i=1}^{N} d(x_i, y_i)$$

Self-Authentication of Documents

§ Constraints:

$$\frac{1}{|\mathcal{K}_{DH}||\mathcal{M}|}\sum_{k_{DH}\in\mathcal{K}_{DH}}\sum_{m\in\mathcal{M}}\sum_{\mathbf{x}\in\mathcal{X}^N} d(\mathbf{x},\psi^N(f^N(m,\mathbf{x},k_{DH}),\mathbf{x}))p_{\mathbf{X}}(\mathbf{x}) \leq D^E$$

$$\sum_{\mathbf{y}\in\mathcal{Y}^N}\sum_{\mathbf{v}\in\mathcal{V}^N} d(\mathbf{y},\mathbf{v})p(\mathbf{v}|\mathbf{y})p(\mathbf{y}) \leq D^A$$

§ Probability of error:

$$P_e^{(N)} = \frac{1}{|\mathcal{K}_{DH}||\mathcal{M}|}\sum_{k_{DH}\in\mathcal{K}_{DH}}\sum_{m\in\mathcal{M}} \Pr\{g(\mathbf{V},k_{DH}) \neq m | K_{DH}=k_{DH}, M=m\}$$

§ Data-hiding capacity for a fixed channel:

$$C = \max_{p(u,w|x)} [I(U;V) - I(U;X)]$$

$U \in \mathcal{U}$
auxiliary r.v.

Self-Authentication of Documents

- Authentication:

$$\eta : \tilde{\mathcal{M}} \times \tilde{\mathcal{M}} \rightarrow \{0,1\}$$

Decision is taken w.r.t. to a predefined threshold.

- Authentication based on the hashing data-hiding separation principle [SPIE2006]:

> If $\mathbf{X}$ is a finite alphabet stochastic process that satisfies the asymptotic equipartition property, then there is a hashing data-hiding scheme with specified probability of authentication error, if the rate of the hashing code R_H satisfies $R_H \leq R_{DH} < C$.

Gel'fand-Pinsker Text Data-Hiding

- Let $\mathbf{X} = (X_1, \ldots, X_N)$ be the text object, where m is to be hidden.
- Each character X_n is a data structure consisting of multiple quantifiable component fields (features): shape (geometric definition), position, orientation, size, color, etc.
- Example:
  - In the so-called Scalar Costa Scheme (SCS) the auxiliary random variable U is approximated by:

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

compensation parameter factor

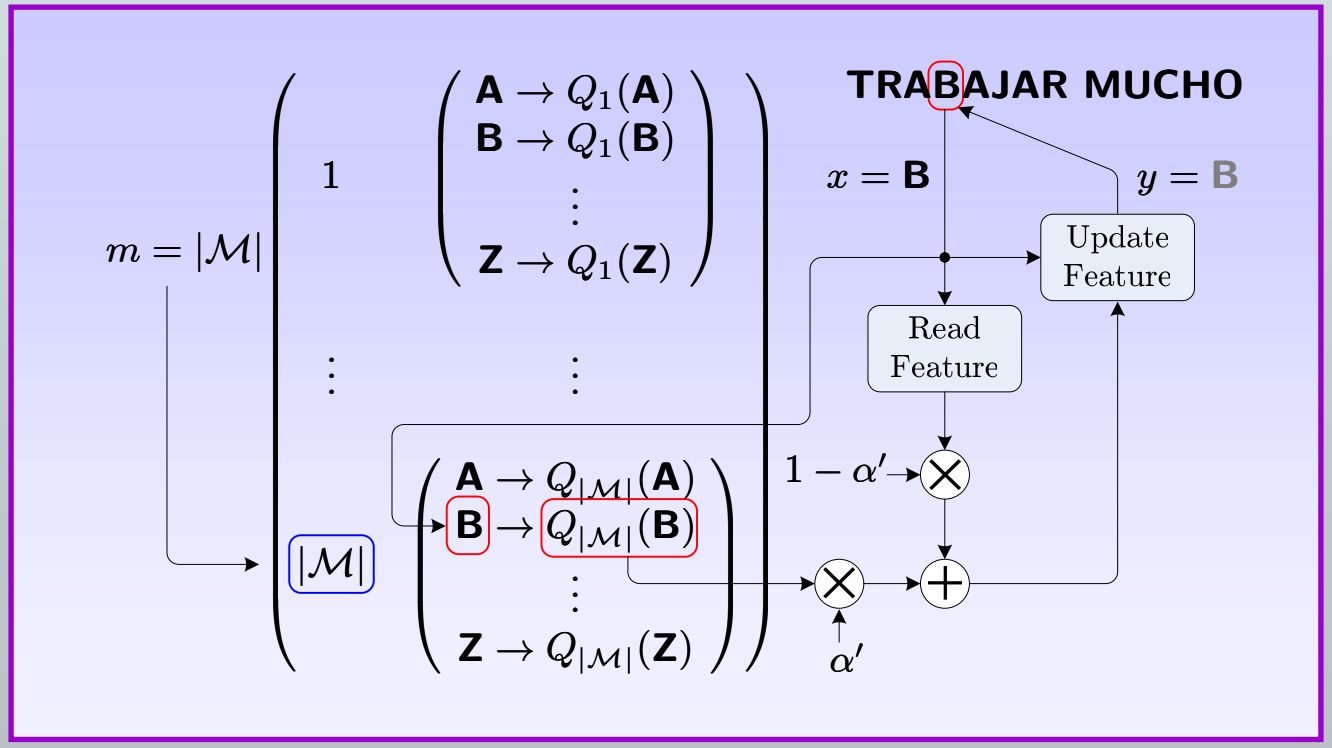
high rate scalar quantizer

  - The resulting stego data is:

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

Gel'fand-Pinsker Text Data-Hiding

- Example (continued):
  - The underlying codebook and encoding mechanism:



Color Index Modulation (CIM)

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

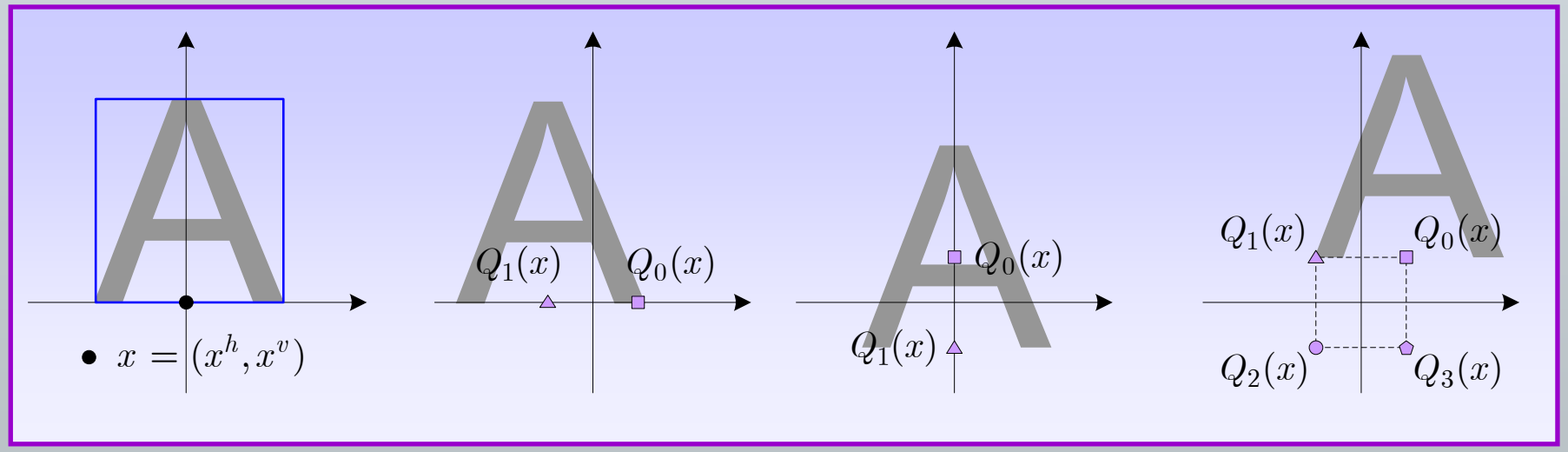
	0 1 0 1 1 0 0 1 0 0 0 1 0 1
VAMOS A TRABAJAR	**VAMOS A TRABAJAR**

- 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.

Note: CIM → printing → halftone index modulation (HIM).

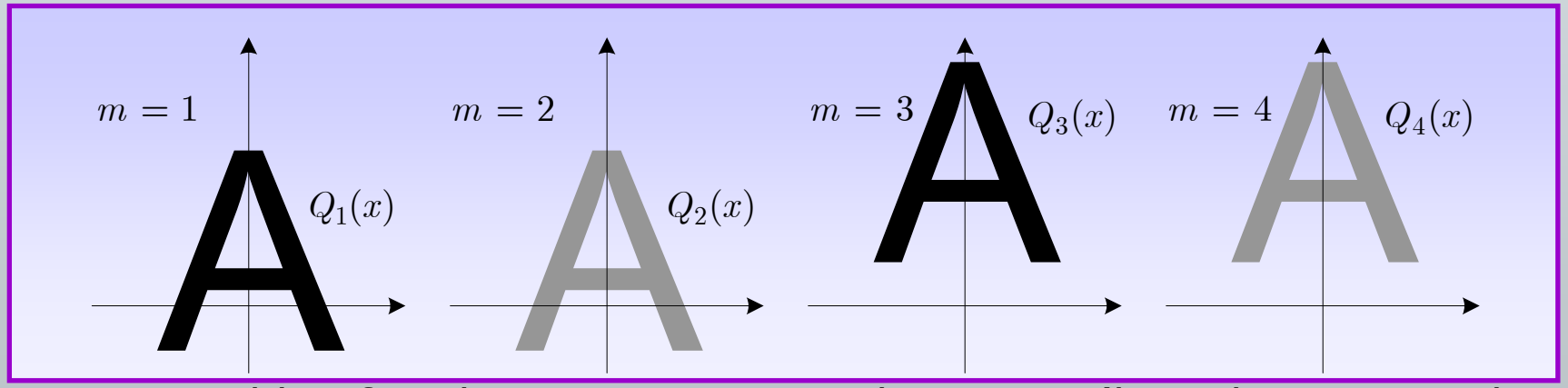
Location Index Modulation (LIM)

- The stego text is obtained via («), where $\alpha' = 1$ and the character feature to quantize is location.
- This method quantizes either the horizontal coordinate X^h, the vertical coordinate X^v, or both.



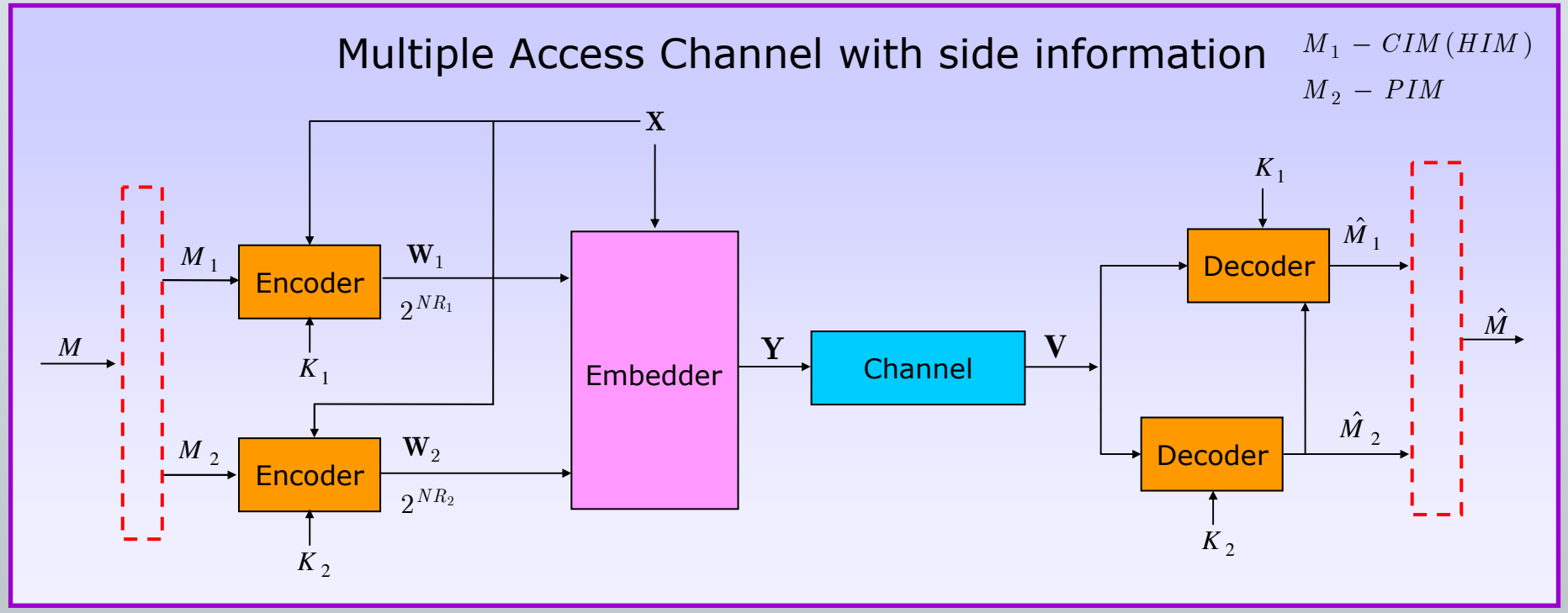
Hybrid Schemes

- Consider simultaneously multiple independent character features instead of a single one.
- Main advantage: higher data storage rate of the resulting scheme.
- Example: CIM + LIM
  - (Raw) data storage rate: 2 bits/character.



  - Capable of authenticating a text document (based on LIM and robust text hashing), and of distinguishing the original from its copies (based on CIM).

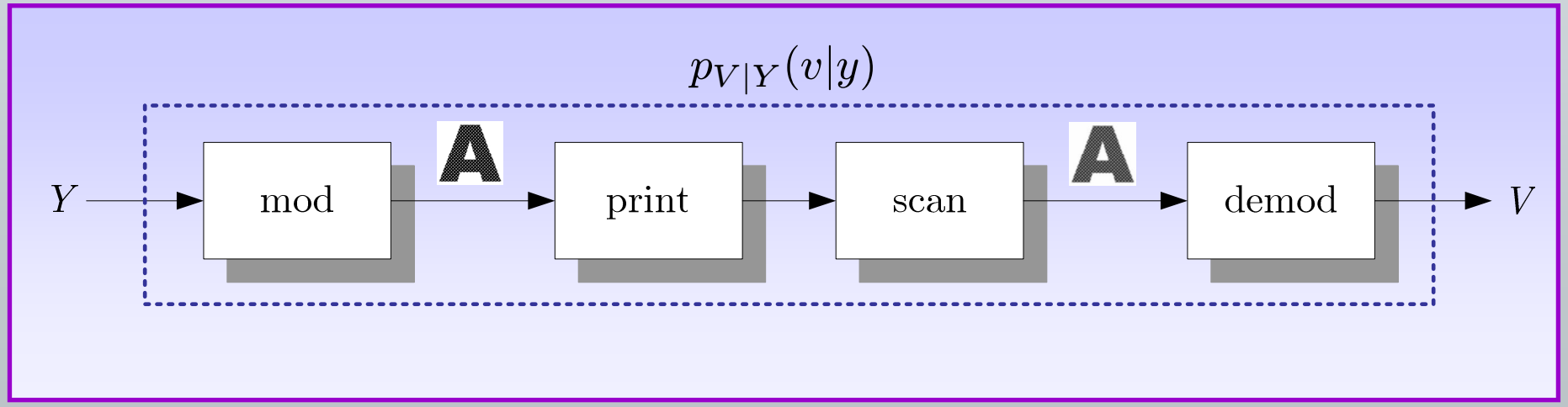
Hybrid Schemes – IT framework



- Rate splitting (R_1, R_2)
- Joint constraint on embedding distortion D^E

Error Control Coding (ECC) for Print-and-Scan Channels

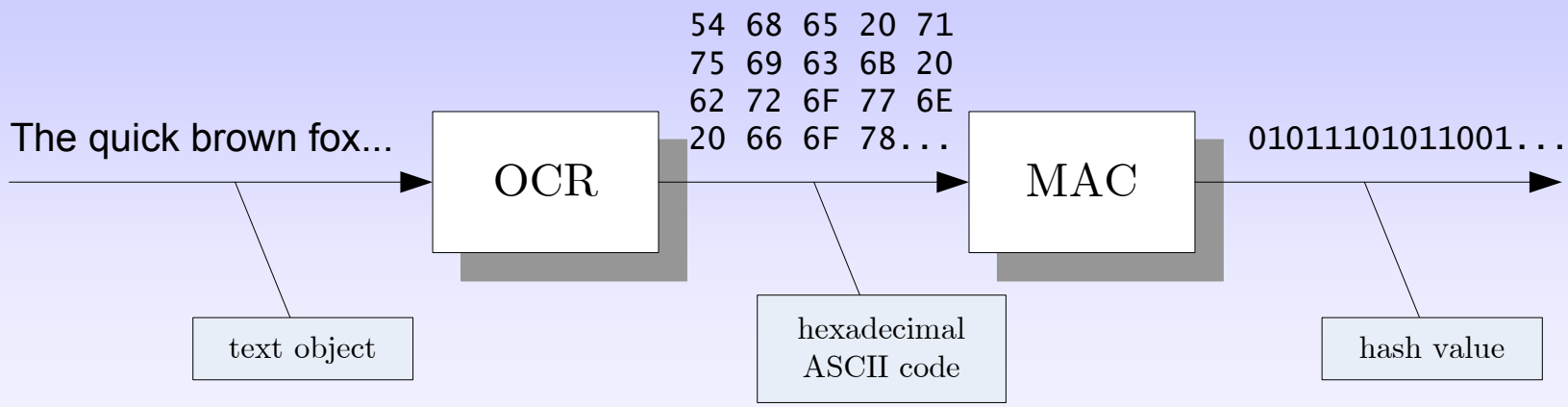
- An outer layer of coding can be used taking into account the *print-and-scan channel*.
- Some modifications to get full benefit of soft-decision decoding techniques.



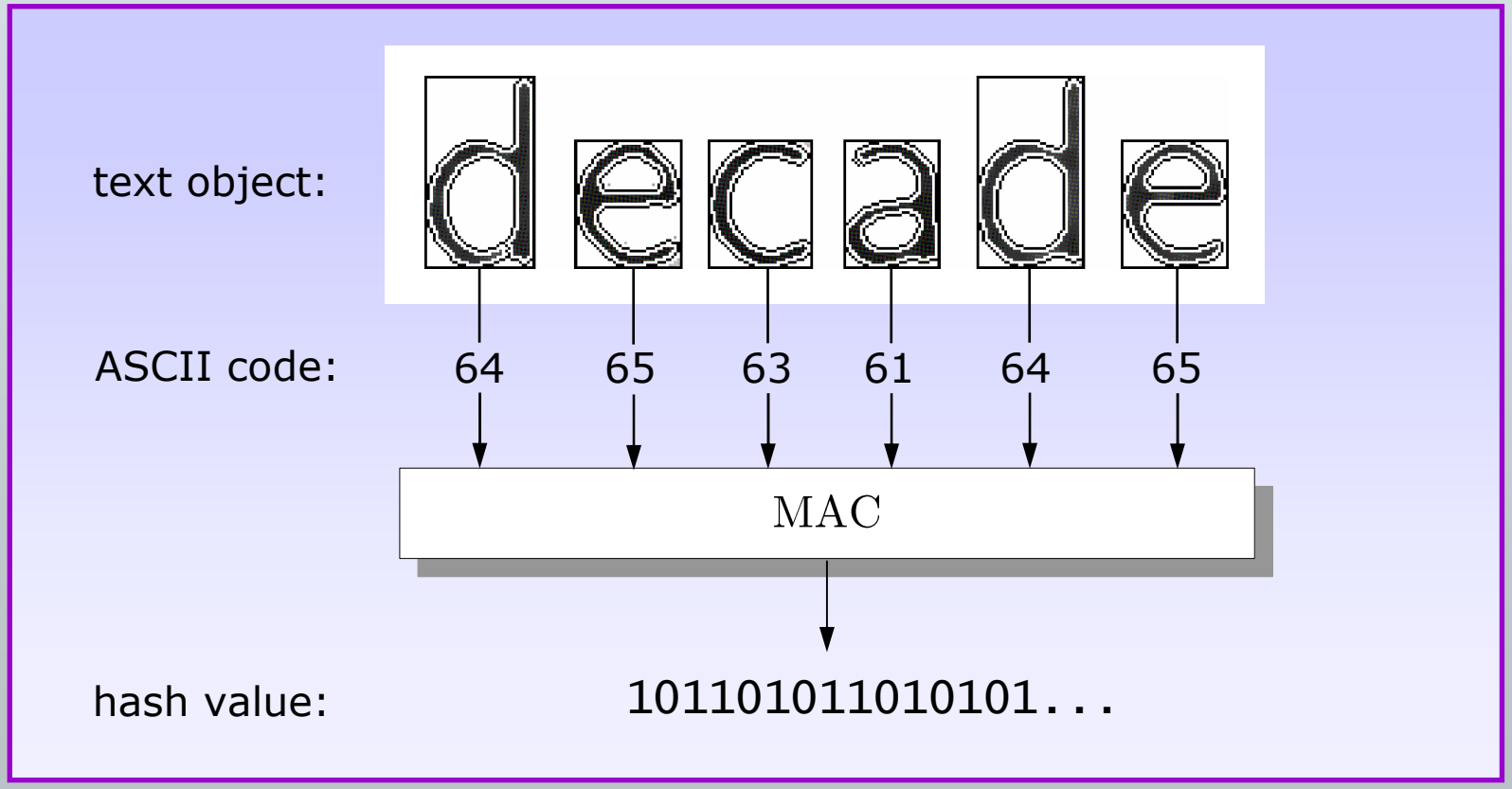
Robust Hashing of Text Documents

- We consider two text hashing techniques.
- The hash value $\mathbf{H} = H(k_H, \mathbf{x})$ is required to be:
  - invariant under legitimate modifications of the text object $\mathbf{x}$ including conversion between electronic formats, data-hiding, printing, scanning, photocopying, faxing, etc.
  - sensitive to illegitimate modifications which change the semantics of $\mathbf{x}$.

OCR + MAC Text Hashing:



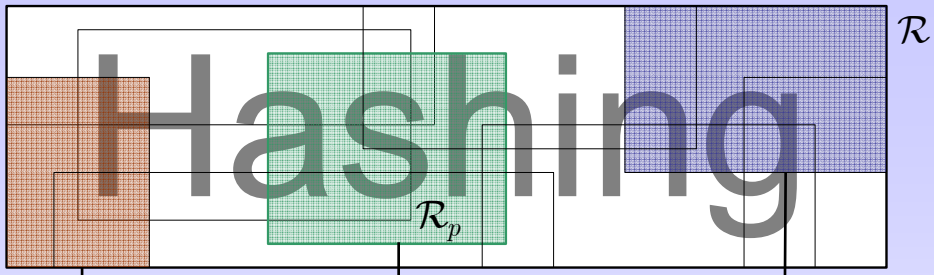
OCR + MAC Text Hashing (continued):



Random Tiling Text Hashing

- Preprocessing: skew correction, segmentation, and bitmap conversion.
- Generate at random P rectangles $\mathcal{R}_p$.

text object:



comparison: $\mu_1 \lessgtr T_1 \quad \ldots \quad \mu_p \lessgtr T_p \quad \ldots \quad \mu_P \lessgtr T_P$

key-dependent threshold

hash value: 1 … 0 … 1

$$\mu_p = \frac{1}{|\mathcal{R}_p| \sum_{r \in \mathcal{R}_p} \lambda_{k_H}(r)} \sum_{r \in \mathcal{R}_p} l(r) \lambda_{k_H}(r)$$

luminance value of pixel located at r

key-dependent weight

Experimental Results

- Standard office equipment was used.
- All digital images containing text lines were created/processed at 600 ppi.
- OCR + MAC text hashing:
  - ABBYY FineReader as OCR tool,
  - HMAC SHA-1 truncated to 80 bits as MAC ($R_H \leq R_{DH}$).
- Random tiling text hashing:
  - $P = 1024$, $\lambda_{k_H}(r) = 1$ for all $r \in \mathcal{R}_p$,
  - The width and height (in pixels) of $\mathcal{R}_p$ drawn uniformly at random from [5,10], [5,50], respectively.
- We used the relative Hamming distance $d_H(\mathbf{h}_1, \mathbf{h}_2)$ to compare any two hash values $\mathbf{h}_1$ and $\mathbf{h}_2$.

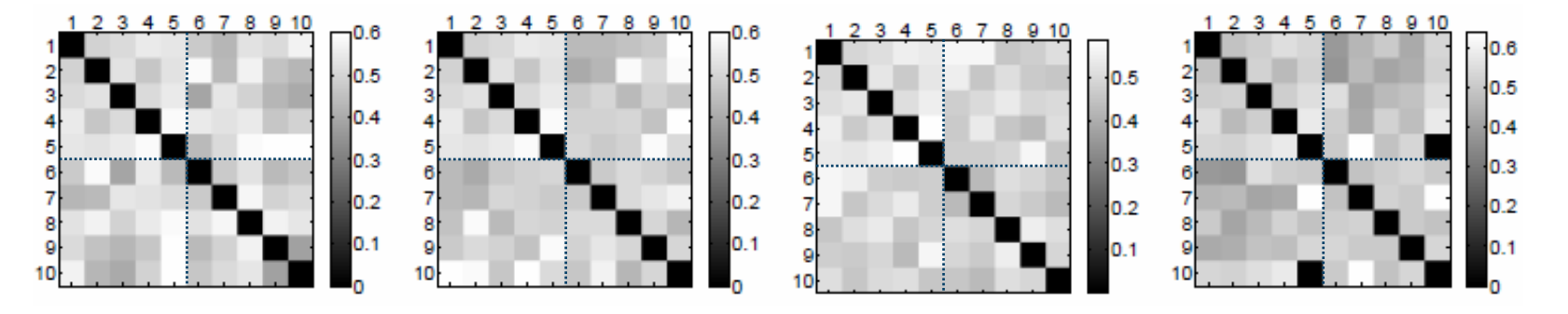
Experimental Results

- Tested **legitimate modifications**:
  - Electronic format conversion (Word ↔ PostScript ↔ PDF), printing and scanning, photocopying, and faxing.
  - Provided the OCR tool does not make a mistake, both text hashing methods show good effectiveness.
- Tested **illegitimate modifications** (digital form):

level → levels	authentication → autentication
(a) addition of one new character	(b) suppression of one character
system → sistem	document → doument
(c) replacement with visually different character	(d) replacement with visually similar character

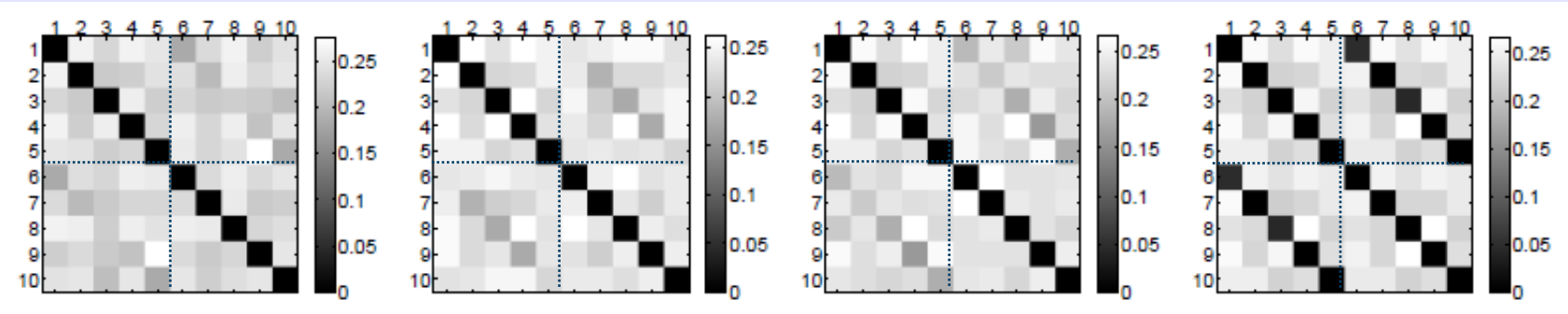
Experimental Results – illegal modifications

§ OCR + MAC text hashing:



(a) addition (b) suppression (c) replacement-d (d) replacement-s

§ Random tiling text hashing:



(a) addition (b) suppression (c) replacement-d (d) replacement-s

Conclusions

- The combination of robust text hashing and text data-hiding is a promising solution to the problem of authentication and tamper-proofing of electronic and printed text documents.
- By combining independent text data-hiding methods (e.g. CIM and LIM) it is possible to increase the data storage rate (IT framework).
- OCR + MAC text hashing shows better applicability than random tiling text hashing. However, OCR + MAC text hashing method highly relies on the accuracy of the OCR tool.

Future research:

- Countermeasures for the weaknesses and security analysis of OCR+MAC text hashing and random tiling text hashing.
- Document authentication with mobile phones.

Conclusions