Authentication of electronic and printed text documents

VILLÁN-SEBASTIÁN, Renato Ficher

Abstract

Electronic and printed text documents are still the most common means of information communication among humans. One of the major problems that our society faces is the ease with which dishonest people can forge official documents. Although modern cryptography can deal with this security issue in the case of electronic documents, it cannot be applied to the case of printed documents. Classical ways to secure a printed document are expensive and usually do not protect the document's content itself. In this thesis, we address the above problems by developing new low-cost methods that can be applied to both electronic and printed text documents. The main contributions include the development of a new state-of-art for two-dimensional bar codes, a novel information theoretic framework for document authentication as well as practical text data hiding and robust text hashing algorithms that can be used as building blocks for text document authentication systems.

Reference


URN : urn:nbn:ch:unige-18237
DOI : 10.13097/archive-ouverte/unige:1823

Available at:
http://archive-ouverte.unige.ch/unige:1823

Disclaimer: layout of this document may differ from the published version.
Authentication of Electronic and Printed Text Documents

THÈSE
présentée à la Faculté des sciences de l’Université de Genève pour obtenir le grade de Docteur ès sciences, mention informatique
par
Renato Fisher VILLÁN SEBASTIÁN
de
La Paz (BOLIVIE)

Thèse N° 3977

GENÈVE
2008
Doctorat ès sciences
mention informatique

Thèse de Monsieur Renato Fisher VILLÁN SEBASTIÁN

intitulée :

"Authentication of Electronic and Printed Text Documents"

La Faculté des sciences, sur le préavis de Messieurs S. VOLOSHYNOVSKIY, professeur adjoint suppléant et directeur de thèse (Département d'informatique), Th. PUN, professeur ordinaire et co-directeur de thèse (Département d'informatique), F. PÉREZ-GONZÁLEZ, professeur (Universidad de Vigo – Departamento de Teoria de la Señal y las Comunicaciones – Vigo, Espagne), F. DEGUILLAUME, Docteur (SICPA Management S.S., – Prilly, Suisse), et O. KOVAL, docteur (Département d'informatique), autorise l'impression de la présente thèse, sans exprimer d'opinion sur les propositions qui y sont énoncées.

Genève, le 20 mai 2008

Le Doyen, Jean-Marc TRISCONE

N.B. - La thèse doit porter la déclaration précédente et remplir les conditions énumérées dans les "Informations relatives aux thèses de doctorat à l'Université de Genève".

Nombre d'exemplaires à livrer par colis séparé à la Faculté : 7
A mis padres
Résumé

Dans cette thèse, nous étudions le problème général d’authentification de documents et, en particulier, celui d’authentification de documents texte. La condition fondamentale imposée aux systèmes étudiés tout au long de ce travail est que ceux-ci doivent fonctionner aussi bien pour la forme électronique que pour la forme imprimée d’un document.

Premièrement, en raison de leur flexibilité, bas prix, et omniprésence, les codes-barres bidimensionnels (2-D) sont étudiés en tant que moyens auxiliaires de stockage de données pour l’authentification de documents. Le problème principal que nous adressons est celui du volume limité de stockage de données des codes-barres 2-D connus concernant certaines applications d’authentification de documents. Nous démontrons, pour la première fois, que les codes-barres 2-D multiniveaux possèdent les taux de stockage de données les plus hauts comparativement aux codes-barres 2-D du domaine publique. Deux contributions principales sont au cœur de ce résultat. La première est la construction d’un modèle simplifié, néanmoins raisonnable, pour le canal impression-numérisation, adapté spécifiquement au problème de communication de données par l’intermédiaire des symboles 2-D multiniveaux. La seconde est l’adaptation du codage multiniveau et décodage multitiéape, technique de modulation codée développée originellement pour le canal additif à bruit blanc Gaussian, au canal impression-numérisation.

Deuxièmement, étant donné que pour quelques applications d’authentification de documents il n’est pas possible d’utiliser des code-barres 2-D, par exemple en raison des exigences esthétiques, nous considérons le filigranage de données en tant que technologie alternative de stockage de données. Dans ce contexte, en utilisant des outils de la théorie de l’information, nous effectuons l’analyse de performance d’un système générique d’authentification de documents ainsi que nous justifions la convenance d’un système d’authentification basé sur le hachage visuel robuste et le filigranage de données. En outre, en utilisant un argument d’entropie conditionnelle, nous conduisons une analyse de sécurité du dernier système qui, contrairement aux études existantes, se fonde crucialement sur les fuites d’information possibles du système d’authentification.
Troisièmement, nous étudions le problème particulier d’authentification de documents texte lorsque l’on utilise le hachage robuste de texte et le filigranage de données textuelles. Dans ce contexte, nous proposons un nouveau cadre théorique pour le problème de filigranage de données textuelles. Nous expliquons comment ce problème peut aussi être vu comme une instance du problème de Gel’fand-Pinsker. L’idée principale est de considérer un caractère de texte en tant que structure de données se composant de multiples attributs quantifiables tels que la forme, la localisation, l’orientation, la taille, la couleur, etc. La puissance de ce cadre théorique est démontrée en prouvant que les techniques précédentes de filigranage de données textuelles, à savoir les méthodes à espace ouvert et les méthodes à attributs de caractère, sont des cas particuliers d’une technique de filigranage de données textuelles basée sur la quantification, laquelle est à son tour une implémentation pratique du système de filigranage de données textuelles selon Gel’fand-Pinsker. À des fins d’illustration, nous présentons la méthode de modulation d’index de couleur en tant que méthode originale pour le filigranage semi-fragile de données textuelles sous formes électronique et imprimé. Le travail expérimental confirme que cette méthode présente une haute invisibilité perceptuelle, un taux de stockage d’information élevé, et est entièrement automatisable. Nous étudions également deux algorithmes de hachage robuste de textes qui sont particulièrement bien adaptés au problème d’authentification de documents texte. Nos résultats confirment que les deux algorithmes de hachage de texte sont robustes contre les dégradations légitimes typiques des documents texte comprenant la conversion électronique de format, l’impression, la numérisation, la photocopie, et l’envoi par fax.

Finalement, nous appliquons nos résultats théoriques et pratiques concernant l’authentification de documents à la conception d’un nouveau système d’authentification pour les documents d’identification biométriques, lequel peut être utilisé avec des dispositifs mobiles tels que les téléphones portables équipés d’un appareil-photo numérique. Nous supposons que le document d’identification biométrique contient des données biométriques sous forme d’image (par exemple, le visage ou l’empreinte digitale) et des données personnelles sous forme de texte, toutes étant directement imprimées sur le document d’identification. La solution proposée utilise le filigranage de données pour stocker les données biométriques hachées sur les données personnelles et vice-versa. Les résultats expérimentaux obtenus prouvent que le système proposé constitue une solution viable et pratique.
Remerciements

Une thèse de doctorat est sans aucun doute un travail de longue haleine qui ne saurait s’achever sans l’aide précieuse de plusieurs personnes.

Je remercie mes superviseurs de thèse, Prof. Sviatoslav Voloshynovskiy et Prof. Thierry Pun, de m’avoir donné l’opportunité de devenir membre du Laboratoire de Vision par Ordinateur et Multimédia (CVML) de l’Université de Genève. Leurs conseils avisés et leur soutien continu se sont avérés inestimables à la conclusion de cette thèse et de ma formation.

J’exprime aussi mon énorme gratitude au Prof. Sviatoslav Voloshynovskiy et au Dr Oleksiy Koval pour avoir guidé et suivi de très près mon travail de recherche. Il est certain que sans leur collaboration et soutien cette thèse n’aurait jamais vu le jour.

Je remercie les membres externes de mon jury de thèse, Prof. Fernando Pérez González et Dr Frédérique Deguillaume, pour le temps consacré à la lecture de ma thèse et leurs précieux commentaires.

Le CVML a été un endroit très convivial et scientifiquement très stimulant. Je remercie les nombreux collègues et amis que j’y ai rencontré au cours de ces dernières quatre années : Dr Stéphane Marchand-Maillet, Dr Eric Bruno, Dr Serhiy Kosinov, Dr Nicolas Moenne-Loccoz, Dr Bruno Janvier, Donn Morrison, Eniko Szekely, Jana Kludas, Dr Teodor Alecu, Dr Julien Kronegg, Dr Karim Ansari-Asl, Dr Romana Rytsar, Guillaume Chanel, Benoît Deville, Mohammad Soleymani, Dr Joep Kierkels, Dr Yuriy Rytsar, Dr Frédérique Deguillaume, Dr José Emilio Vila Forcén, Dr Emre Topak et Fokko Beekhof. Bien que rattachés à d’autres laboratoires, je remercie par la même occasion Katarzyna Wac, Luminita Moraru et Jee-Hyub Kim pour leur amitié et convivialité.

Je remercie aussi le personnel administratif et technique du Centre Universitaire d’Informatique (CUI), en particulier, Germaine Gusthiot, Catia Bertossa, Daniel Agulleiro, Nicolas Mayencourt, Dorothée Hauser et Amélia Bossard pour avoir facilité ma vie quotidienne au sein du CUI.

Finalement, je remercie infiniment toutes les personnes qui m’ont toujours soutenu avec leur amour et encouragement : ma mère, Delia, Silvia, Maritza, Jhonny, Omar, mes amis les plus proches et Nathalie.
Contents

List of Figures xiii
List of Tables xvii
Notations xix

1 Introduction 1
1.1 Hash-Based Document Authentication .............................. 1
1.2 Thesis Scope ...................................................... 4
1.3 Main Contributions ................................................ 6
1.4 Thesis Outline ...................................................... 8

2 Multilevel 2-D Bar Codes 11
2.1 Introduction .......................................................... 11
2.2 Framework and Theoretical Performance Limits .................. 13
2.3 Discrete Print-Scan Channel Model ................................ 14
   2.3.1 Characterization of the Discrete Print-Scan Channel .......... 15
   2.3.2 Characterization of the Noise Probability Distribution ....... 19
   2.3.3 Discrete Print-Scan Channel Model .............................. 21
2.4 Multilevel Coding for the Print-Scan Channel .................... 23
   2.4.1 MLC/MSD for the AWGN Channel ............................... 23
   2.4.2 MLC/MSD for the Print-Scan Channel ........................... 28
2.5 Experimental Results on the Design of Multilevel 2-D Bar Codes . 31
   2.5.1 Constellation and MLC/MSD Rate Design ....................... 31
   2.5.2 Performance Evaluation of Multilevel 2-D Bar Codes .......... 31
   2.5.3 Fine Tuning of Multilevel 2-D Bar Codes ...................... 33
2.6 Conclusions and Related Work ...................................... 37

3 Information-Theoretic Framework for Document Authentication 39
3.1 Introduction .......................................................... 39
3.2 The General Document Authentication Problem .................... 40
<table>
<thead>
<tr>
<th>Section</th>
<th>Title</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>3.3</td>
<td>Authentication via Robust Hashing and Data Hiding</td>
<td>43</td>
</tr>
<tr>
<td>3.4</td>
<td>Security Analysis of Document Authentication</td>
<td>52</td>
</tr>
<tr>
<td>3.4.1</td>
<td>The Case of Identical Keys</td>
<td>53</td>
</tr>
<tr>
<td>3.4.2</td>
<td>The Case of Different Keys</td>
<td>54</td>
</tr>
<tr>
<td>3.5</td>
<td>Conclusions</td>
<td>55</td>
</tr>
<tr>
<td>4</td>
<td>Practical Methods for Text Document Authentication</td>
<td>57</td>
</tr>
<tr>
<td>4.1</td>
<td>Introduction</td>
<td>57</td>
</tr>
<tr>
<td>4.2</td>
<td>State of the Art of Text Data Hiding</td>
<td>58</td>
</tr>
<tr>
<td>4.3</td>
<td>Text Data-Hiding as a Gel’fand-Pinsker Problem</td>
<td>58</td>
</tr>
<tr>
<td>4.3.1</td>
<td>Costa’s Problem</td>
<td>59</td>
</tr>
<tr>
<td>4.3.2</td>
<td>Practical Implementation of Costa’s Problem</td>
<td>59</td>
</tr>
<tr>
<td>4.4</td>
<td>Practical Implementation of Gel’fand-Pinsker Text Data-Hiding</td>
<td>61</td>
</tr>
<tr>
<td>4.4.1</td>
<td>Color Index Modulation</td>
<td>61</td>
</tr>
<tr>
<td>4.4.2</td>
<td>Location Index Modulation</td>
<td>63</td>
</tr>
<tr>
<td>4.4.3</td>
<td>Halftone Index Modulation</td>
<td>64</td>
</tr>
<tr>
<td>4.4.4</td>
<td>Hybrid Schemes</td>
<td>65</td>
</tr>
<tr>
<td>4.4.5</td>
<td>Error Control Coding for Print-Scan Channels</td>
<td>66</td>
</tr>
<tr>
<td>4.5</td>
<td>Robust Hashing of Text Documents</td>
<td>68</td>
</tr>
<tr>
<td>4.5.1</td>
<td>OCR + MAC Text Hashing</td>
<td>68</td>
</tr>
<tr>
<td>4.5.2</td>
<td>Random Tiling Text Hashing</td>
<td>68</td>
</tr>
<tr>
<td>4.6</td>
<td>Experimental Results</td>
<td>70</td>
</tr>
<tr>
<td>4.6.1</td>
<td>Text Data Hiding</td>
<td>70</td>
</tr>
<tr>
<td>4.6.2</td>
<td>Robust Text Hashing</td>
<td>71</td>
</tr>
<tr>
<td>4.7</td>
<td>Conclusions</td>
<td>75</td>
</tr>
<tr>
<td>5</td>
<td>Authentication of Biometric Identification Documents</td>
<td>77</td>
</tr>
<tr>
<td>5.1</td>
<td>Introduction</td>
<td>77</td>
</tr>
<tr>
<td>5.2</td>
<td>Identification Framework</td>
<td>78</td>
</tr>
<tr>
<td>5.2.1</td>
<td>Biometrics</td>
<td>78</td>
</tr>
<tr>
<td>5.2.2</td>
<td>Personal Data</td>
<td>78</td>
</tr>
<tr>
<td>5.2.3</td>
<td>Document-Based Identification</td>
<td>79</td>
</tr>
<tr>
<td>5.3</td>
<td>Requirements and Proposed Approach</td>
<td>80</td>
</tr>
<tr>
<td>5.3.1</td>
<td>Mobile Device Architecture</td>
<td>80</td>
</tr>
<tr>
<td>5.3.2</td>
<td>Identity Document Design</td>
<td>80</td>
</tr>
<tr>
<td>5.4</td>
<td>Problem Formulation and Solution Approaches</td>
<td>83</td>
</tr>
<tr>
<td>5.4.1</td>
<td>Authentication of Biometric Identification Documents</td>
<td>83</td>
</tr>
<tr>
<td>5.4.2</td>
<td>Authentication Using Cross-Storage of Hashes</td>
<td>84</td>
</tr>
<tr>
<td>5.5</td>
<td>Practical Scheme for Document Authentication</td>
<td>87</td>
</tr>
<tr>
<td>5.5.1</td>
<td>Information Storage with Digital Data Hiding</td>
<td>87</td>
</tr>
<tr>
<td>5.5.2</td>
<td>Robust Visual Hashing</td>
<td>87</td>
</tr>
<tr>
<td>5.5.2.1</td>
<td>Robust Visual Hashing for Personal Data</td>
<td>87</td>
</tr>
<tr>
<td>5.5.2.2</td>
<td>Robust Visual Hashing for Biometrics</td>
<td>88</td>
</tr>
<tr>
<td>5.6</td>
<td>Experimental Results</td>
<td>89</td>
</tr>
<tr>
<td>5.7</td>
<td>Conclusions</td>
<td>93</td>
</tr>
</tbody>
</table>
6 Conclusions and Future Research Directions 95
  6.1 Conclusions ................................................................. 95
  6.2 Future Research Directions ............................................. 96

Bibliography 99

Publications by Renato F. Villán S. 107

Acronyms and Abbreviations 109

List of Symbols 113
List of Figures

1.1 Document authentication system that stores hashes in a database: (a) document tamper protection, (b) document authentication. ............................ 3
1.2 Different examples of on-document data storage means: (a) 2-D bar code, (b) magnetic stripe, (c) RFID tag, (d) electronic memory chip, (e) CD, (f) optical memory. ................................. 4
1.3 Document authentication system based on hash storage in a dense 2-D bar code: (a) document tamper protection, (b) document authentication. ............................... 5
1.4 Document authentication system based on hash storage in a sparse 2-D bar code: (a) document tamper protection, (b) document authentication. ............................... 6
1.5 Document authentication system based on hash storage using data hiding: (a) document tamper protection, (b) document authentication. ............................... 7

2.1 (a) Square halftone cell of \(a\)-by-\(a\) dots. (b) Array of halftone cells. ............................... 13
2.2 Maximum number \(U\) of bits per square inch for a multilevel 2-D bar code in the ideal setup. ............................... 14
2.3 Modulation and demodulation steps for communications over the P-S channel. ............................... 15
2.4 Multilevel 2-D bar code. (a) Original digital image: \(2 \times 2\) pixel symbols, 1 pixel of inter-symbol space, \(r_{im} = 200\) ppi. (b) Printed-and-scanned digital image: \(r_p = 600\) dpi, \(r_s = 600\) ppi, \(6 \times 6\) pixel noisy symbols. ............................... 16
2.5 Sample mean \(\hat{\mu}_{Y|X}(x)\) for different combinations of printers and scanners. ............................... 17
2.6 Sample standard deviation \(\hat{\sigma}_{Y|X}(x)\) for different combinations of printers and scanners. ............................... 18
2.7 Normalized histogram of the received signal points and GGD approximation for each \(x \in \mathcal{X}\). ............................... 20
2.8 Discrete model for the P-S channel. ............................... 21
2.9 Transmission of an 8–PAM equidistant constellation over the AWGN channel. ............................... 24
2.10 Conditional p.d.f.’s for the equivalent channel at level 0 when an 8–PAM equidistant constellation with Ungerböck’s labeling is used. ............................... 25
2.11 Conditional p.d.f.’s for the equivalent channel at level 1 when an 8–PAM equidistant constellation with Ungerböck’s labeling is used. 25
2.12 Multilevel encoder. 26
2.13 Multistage decoder. 26
2.14 Capacity rule for selecting the individual code rates. 27
2.15 Conditional p.d.f.’s for the equivalent channel at level 0 when an 8–PAM non-equidistant constellation with Ungerböck’s labeling is used for communications over a P-S channel. 29
2.16 Conditional p.d.f.’s for the equivalent channel at level 1 when an 8–PAM non-equidistant constellation with Ungerböck’s labeling is used for communications over a P-S channel. 29
2.17 Building up the non-equidistant 8–PAM signal constellation X for the \( p_2 = s_3 \) combination, \( k = 1.09 \). 30
2.18 Modulation parameters for multilevel 2-D bar codes. (a) \( 2 \times 2 \) pixel 2-D symbols, 1 pixel of inter-symbol space, \( r_{im} = 200 \) ppi. (b) \( 6 \times 6 \) pixel 2-D symbols, 1 pixel of inter-symbol space, \( r_{im} = 600 \) ppi. 32
3.1 Block diagram of a generic document authentication system. 40
3.2 Document authentication based on the separation of robust hashing and data hiding. 44
3.3 Discrete memoryless data hiding channel. 45
3.4 Gel’fand-Pinsker codebook construction. 48
3.5 Gel’fand-Pinsker encoder. 48
3.6 Gel’fand-Pinsker decoder. 49
3.7 Generic approach to robust hashing. 49
4.1 Text data hiding as an instance of the Gel’fand-Pinsker problem. 58
4.2 SCS codebook for text data hiding (\( N = 1 \)). 60
4.3 Color Index Modulation: (a) original text; (b) marked text (exaggerated). 62
4.4 Two-level color index modulation: (a) original text; (b) marked text using \( Q_0(x) = 0 \) and \( Q_1(x) = 46 \). 63
4.5 Location Index Modulation (\( N = 1 \)). 64
4.6 Location Index Modulation (\( N = 5 \)). 65
4.7 Halftone index modulation: (a) original character; (b) marked character for \( m = 0 \); (c) marked character for \( m = 1 \). 66
4.8 Channel model for electronic-printed text document authentication. 67
4.9 OCR + MAC text hashing. 68
4.10 Random tiling text hashing: (a) character-based, \( x = W \); (b) word-based, \( x^T(H, a, s, h, i, n, g) \). 69
4.11 Segmentation of characters. 71
4.12 Sample text lines. 72
4.13 Sample of forged text lines: (a) by adding one new character; (c) by suppressing one character; (c) by replacing one character with a visually different one; (d) by replacing one character with a visually similar one. 73
4.14 Normalized Hamming distances for legitimate distortions and OCR + MAC text hashing: (a) electronic format conversion; (b) printing and scanning; (c) photocopying; (d) faxing. .......................................................... 73
4.15 Normalized Hamming distances for legitimate distortions and random tiling text hashing: (a) electronic format conversion; (b) printing and scanning; (c) photocopying; (d) faxing. ................................. 74
4.16 Normalized Hamming distances for illegitimate distortions and OCR + MAC text hashing: (a) addition of one new character; (c) suppression of one character; (c) replacement of one character by a visually different one; (d) replacement of one character by a visually similar one. ...... 74
4.17 Normalized Hamming distances for illegitimate distortions and random tiling text hashing: (a) addition of one new character; (c) suppression of one character; (c) replacement of one character by a visually different one; (d) replacement of one character by a visually similar one. ....... 75

5.1 The relations between an authority, a person, a challenger and an identification document. ............................................................... 79
5.2 Proposed system architecture for document authentication. .............. 81
5.3 Variants of identification documents with a secure link between biometric data and personal data: (a) via a database; (b) via an on-document storage module; (c) via cross-storage using robust hashing and data hiding. .... 81
5.4 Block diagram of a generic biometric document authentication system. .. 83
5.5 Proposed enrollment and production of an identification document. .... 85
5.6 Proposed authentication of an identification document. ................. 86
5.7 Identification document sample. ............................................. 89
5.8 Hamming distances between different images. ............................ 91
5.9 Hamming distances between 64-bit hashes of images before and after being watermarked, printed and scanned, which are legitimate distortions. The Hamming distances of different images are provided for reference. 92
List of Tables

2.1 Printers and scanners used for experimentation. .................. 16
2.2 Tested signal constellations for characterizing the noise probability distribution. .................................................. 19
2.3 Estimated parameters for the GGD approximation of the noise probability distribution. ............................................ 22
2.4 Individual code rates and total rate for multilevel 2-D bar codes using MLC/MSD ($N \to \infty$). ...................................... 32
2.5 BER results of multilevel 2-D bar codes for $R = 2$ ($N = 2048$). .... 34
2.6 Selected parameters for multilevel coding ($N = 2048$). ............... 35
2.7 BER results of multilevel 2-D bar codes ($N = 2048$). ................ 36
2.8 High rate 2-D bar codes. .................................................... 37
4.1 Equipment used for experimentation. ................................. 70
4.2 Performance of two-level color index modulation ($J = 4104$). ........ 72
In this thesis, we use small or capital letters for constants, deterministic variables, and function names. We also use capital letters, e.g. $X$, to denote scalar random variables, bold capital letters, e.g. $X$, to denote vector random variables, and corresponding small letters, e.g. $x$ and $x$, to denote their realizations. The superscript $N$ is used to designate length–$N$ vectors, e.g. $X = X^N = (X_1, X_2, \ldots, X_N)$.

Calligraphic letters, e.g. $\mathcal{X}$, denote sets and $|\mathcal{X}|$ denotes the cardinality of $\mathcal{X}$. The set of all real numbers is denoted by $\mathbb{R}$ and the set of all nonnegative real numbers by $\mathbb{R}^+$. The set $\{0, 1, \ldots, 255\}$ is denoted by $[0, 255]$.

The probability mass (respectively, density) function (p.m.f.) (respectively, p.d.f.) of a discrete (respectively, continuous) random variable $X$ is denoted by $p_X$ (respectively, $f_X$). When no confusion is possible we only write $p(x)$ (respectively, $f(x)$) instead of $p_X(x)$ (respectively, $f_X(x)$).

We use $X \sim p_X(x)$ to indicate that the random variable $X$ is distributed according to $p_X$. The mathematical expectation of a random variable $X \sim p_X(x)$ is denoted by $E_{p_X}[X]$ or simply by $E[X]$ or $\mu_X$. $\text{Var}[X]$ or $\sigma_X^2$ denote the variance of $X$. We employ $X \sim \mathcal{N}(\mu, \sigma^2)$ to indicate that $X$ is distributed according to a Gaussian distribution with mean $\mu$ and variance $\sigma^2$. We use $X \sim \mathcal{GGD}(\mu, \sigma, \gamma)$ to indicate that $X$ is distributed according to a generalized Gaussian distribution with mean $\mu$, standard deviation $\sigma$, and shape parameter $\gamma$.

The entropy (respectively, differential entropy) of a discrete (respectively, continuous) random variable $X$ is denoted by $H(X)$ (respectively, $h(X)$) and the mutual information between random variables $X$ and $Y$ by $I(X; Y)$. The conditional relative entropy between $p_{Y|X}^0$ and $p_{Y|X}^1$ is denoted by $D(p_{Y|X}^1 \parallel p_{Y|X}^0)$.

The length–$S$ binary representation of an index $L$ is denoted by $B^S(L) = (B_1(L), B_2(L), \ldots, B_S(L))$. 

Notations
1 Introduction

TODAY, TEXT DOCUMENTS are still the most common and almost unavoidable means of information communication among humans. Text documents are omnipresent on one’s everyday life in the form of newspapers, books, web pages, contracts, advertisements, official documents, identification documents, etc. Moreover, they can be widely distributed in electronic form via computer networks such as the Internet.

The wide significance of text documents justifies the importance of their copyright protection and authentication, which still remain open and challenging problems. From the point of view of the data hiding community, one possible explanation of the current situation is that text media have a relatively small number of features that can be exploited in order to hide (or embed) information in comparison to images, audio or video. For example, while it is often possible to perform imperceptible modifications to an image, casual readers can easily notice extra letters or punctuation symbols in a text. Indeed, a text document can be seen as a kind of a highly structured image, which is precisely the kind of images the human visual system is more sensitive to. For this reason, the data embedding rate in text media is comparatively much smaller than in images, audio or video. Moreover, the hidden data can always be removed by using optical character recognition (OCR). Therefore, the copyright protection of text documents solely based on robust data hiding or watermarking seems to be questionable. Fortunately, in many practical situations, one is more interested in the authentication of text information rather than in its copyright protection (the exception being the copyright protection of printed papers and books). Thus, the main goal of this thesis is to investigate efficient low-cost technologies for text document authentication.

1.1 Hash-Based Document Authentication

A document authentication system aims at deciding whether a given document is authentic or not. We assume that a document can include text, images, logos, etc. The decision about authenticity is performed at the global level, meaning that the system gives only a binary answer about the entire document: authentic or fake. Contrarily, if the system makes decisions at the local level (paragraph level, line level, image level, subimage level, etc.), we refer to it as a document tamper-proofing system. Thus, document authentication
and tamper-proofing aim at verifying the authenticity of a document and, if the document is found to be fake, at indicating the location of the local modifications.

One possible solution to the document authentication problem consists in the generation of the document’s hashed value based on the knowledge of a secret key $K_H$ [1]. The hashed value, or simply hash, is computed from the entire document and is securely stored somewhere. For authentication, the hash is computed again from the document under investigation and compared with the one that was previously stored. In the case of an authentic document, the two hashes should be identical or, at least, very close to each other in some metric sense. If they are not, the authentication system should declare the document to be a fake. Obviously, the hash function that produces hashes should be designed to withstand various intentional/unintentional legitimate modifications that might occur during the document’s life cycle. At the same time, the hash function should be sensitive enough to various intentional malicious modifications. In image processing, the development of such hash functions is an active field of research known as robust visual hashing [2]. Contrary to document authentication, where the hash is computed from the entire document, document tamper-proofing is based on the concept of local hashing. The document is divided into a number of parts and a local hash is computed for each part. The local hashes are again securely stored somewhere. In this manner, if the document is maliciously modified, the tamper-proofing system is able to identify those local parts where the modifications were introduced. This characteristic is used to provide the user with some hints and evidence about the introduced modifications.

Basically, there exist three approaches to hash-based document authentication and, by extension, to hash-based document tamper-proofing, depending on where the global or local hashes are stored. These are:

- hash storage in an electronic database,
- hash storage onto the document itself using auxiliary data storage means such as two-dimensional (2-D) bar codes, and
- hash storage into the document’s content itself, also known as self-authentication, using data hiding techniques.

A document authentication system that stores hashes in an electronic database is shown in Figure 1.1. At the document tamper protection stage, the hash is computed from the entire document which can include text, images, logos, drawings, etc., and then stored in the hash database. At the document authentication stage, it is only needed to compare the hash computed from the document under investigation to the one stored in the hash database. The drawback of this approach is that in order to authenticate a document there needs to be a direct communication link with the hash database, which might represent an important practical limitation. Moreover the hash database may become a critical bottleneck if it suffers failures and denial-of-service attacks [3].

To overcome the above problems, the document’s hash can be stored directly onto the document itself using auxiliary data storage means. For instance (see Figure 1.2), one can use specific 2-D bar codes, magnetic stripes, radio frequency identification (RFID) tags, electronic memory chips, compact discs (CDs), optical memories, etc., in order to store the hash onto the document [4]. However, the most convenient and cheapest solution
1.1 Hash-Based Document Authentication

is to use a 2-D bar code that can be either directly integrated into the digital document (thus printed in a single step together with the document), or printed on top of an existing physical document. In such a system, as an additional security feature, the hash may be encrypted using a secret key $K_{BC}$ in order to make it more difficult for a forger to break the robust hashing subsystem.

One can distinguish two types of 2-D bar codes, namely dense and sparse 2-D bar codes. Dense 2-D bar codes are usually printed in a localized position, for example at the document’s bottom left margin. Alternatively, one can use sparse 2-D bar codes which are usually printed along the whole document’s surface, potentially in an invisible way using special inks or crystals or even specially chosen normal inks. The corresponding document authentication systems are shown in Figures 1.3 and 1.4.

Although the second approach to document authentication resolves the open issues of the first approach, it has some disadvantages in terms of esthetics if dense 2-D bar codes are used, and flexibility in handling different electronic formats if sparse 2-D bar codes are used. Additionally, the data storage rates of sparse 2-D bar codes subject to visibility constraints are significantly lower than those of dense 2-D bar codes, which is an important factor influencing the authentication power of the robust hashing subsystem.

Finally, the self-authentication approach that falls into the category of digital data
hiding [5] seems to be a very attractive solution for various reasons. First, the document’s hash is securely stored into the document’s content itself based on a secret key $K_{DH}$. Document authentication is performed directly based on the document under investigation without accessing a hash database, similarly to the 2-D bar code authentication approach. Nonetheless, contrary to this approach, the tamper-protected documents remain visually undistinguishable from the original versions. Moreover, as will be shown in Chapters 4 and 5, this approach can be easily integrated into any text/image editing tool and the protected document can be stored in any suitable electronic format or even re-converted from one format to another. Last, it should also be mentioned that the storage data rates of data hiding techniques might be higher than those of sparse 2-D bar codes but lower than those of dense 2-D bar codes. A document authentication system based on data hiding is depicted in Figure 1.5.

1.2 Thesis Scope

In the scope of this thesis, we investigate the general problem of document authentication and in particular of text document authentication. The fundamental requirement imposed on the systems studied in this thesis is that they should apply to both the electronic and printed forms of a document.
Due to their flexibility, low price and omnipresence, 2-D bar codes are studied as auxiliary data storage means for hash-based document authentication. The main targeted problem is that of limited data storage capability of current 2-D bar codes for some document authentication applications. As a solution to this problem, multilevel 2-D bar codes are developed and fully investigated.

Being aware that for some document authentication applications it is not possible to use a 2-D bar code, for example because of esthetical reasons, we consider data hiding as an alternative data storage technology for hash-based document authentication. In this context, we carry out an information-theoretic performance analysis of a generic document authentication system as well as justify the appropriateness of a document authentication system based on robust visual hashing and data hiding. Furthermore, by using an information equivocation argument, we also conduct a security analysis of the latter and propose brute force attacking strategies.

We next tackle the particular problem of text document authentication via robust text hashing and text data hiding. We propose a novel theoretical framework for the problem
of data hiding in text documents and study two robust text hashing algorithms, namely OCR + MAC text hashing and random tiling text hashing, that are particularly well suited for the considered authentication problem.

Finally, we apply our theoretical and practical findings on (text) document authentication in order to design a novel authentication system for biometric identification documents, which may be used in conjunction with mobile devices such as mobile phones or personal digital assistants (PDAs) equipped with a digital camera.

### 1.3 Main Contributions

The main contributions of this thesis can be summarized as follows:

- A novel discrete model for the print-scan (P-S) channel is proposed. This model assumes no inter-symbol interference (ISI) and perfect synchronization, but it takes
into account the fundamental dependence between the channel input and the noise. The main applications of this model are found in the design of high rate multilevel 2-D bar codes and of color index modulation-based text data hiding codes.

- multilevel coding (MLC) / multistage decoding (MSD), a coded modulation technique originally developed for the additive white Gaussian noise channel, is adapted to the P-S channel. It is shown that even though the noise depends on the channel input, the maximum achievable rate of any luminance modulation system can be approached using MLC/MSD.

- For the first time, a parameter set for multilevel 2-D bar codes is given that leads to the highest data storage rates in comparison to other public domain 2-D.

- An information-theoretic performance analysis of a generic document authentica-
tion system is carried out as well as a justification of the appropriateness of a document authentication system based on robust visual hashing and data hiding.

• A novel theoretical framework for the problem of data hiding in text documents is proposed. We explain how this problem can also be seen as an instance of the well-known Gel’fand-Pinsker (G-P) problem. The power of this framework is demonstrated by showing that previous text data hiding techniques, namely open space methods and character feature methods, are particular cases of a general quantization-based text data hiding scheme, which in turn is a practical implementation of the general G-P text data hiding scheme.

• An original G-P text data hiding method, known as color index modulation, for semi-fragile data hiding of electronic and printed text documents is developed. The experimental work confirms that this method has high perceptual invisibility, high information embedding rate, and is fully automatable.

• A novel authentication system for biometric identification documents is proposed. This system may be used in conjunction with mobile devices such as mobile phones or PDAs equipped with a digital camera.

1.4 Thesis Outline

In Chapter 2, we deal with the design of high-rate multilevel 2-D bar codes for the P-S channel. Firstly, we introduce a framework for evaluating the performance limits of these codes by studying an ISI-free, synchronous, and noiseless P-S channel, where the input and output alphabets are finite and the printer device uses halftoning to simulate multiple gray levels. Secondly, we present a new model for the P-S channel specifically adapted to the problem of communications via multilevel 2-D bar codes. This model, inspired by our experimental work, assumes no ISI and perfect synchronization, but independence between the channel input and the noise is not supposed. We adapt the theory of MLC/MSD to the P-S channel. Finally, we present experimental results concerning the design of multilevel 2-D bar codes using MLC/MSD.

In Chapter 3, we consider the general problem of document authentication for the most typical practical scenarios. We formulate an information-theoretic framework for this problem as well as analyze its performance limits. We also justify the appropriateness of a document authentication system based on robust visual hashing and data hiding. For this system, by using an information equivocation argument, we further conduct a security analysis proposing brute force attacking strategies with corresponding complexity estimates.

In Chapter 4, we deal with the problem of authentication of text documents that can be distributed in electronic or printed forms. We advocate the combination of robust text hashing and text data hiding technologies as an efficient solution to this problem. First, we consider the problem of text data hiding. We introduce and explain the G-P text data hiding framework. Second, we study two approaches to robust text hashing that are well suited for the considered problem. Last, the experimental work presents results for a selected G-P text data hiding method and both text hashing approaches.
In Chapter 5, we apply our theoretical and practical findings on (text) document authentication in order to design a novel authentication system for biometric identification documents, which may be used in conjunction with mobile devices equipped with a digital camera. We explain how this solution makes use of data hiding in order to cross-store the robust hashed biometric data inside the personal data and vice versa. Finally we present the relevant experimental results concerning the proposed system.

Finally, in Chapter 6, we draw the conclusions of this work and propose future research directions.
Multilevel 2-D Bar Codes

2.1 Introduction

Two-dimensional (2-D) bar codes are at present widely used in various applications due to their numerous advantages over alternative technologies. Besides being cheap and simple, their main advantage is that they can carry a significant amount of information on surfaces such as paper, plastic, or even ceramics. 2-D bar codes can be either directly printed on the above surfaces or engraved using appropriate lasers [6, 7]. Moreover, printing can be performed using either visible, ultraviolet, or infrared inks, depending on the application’s concern about security. 2-D bar codes can be read using low-resolution readers equipped with cheap charged coupled devices (CCDs) like those in flatbed scanners, handy scanners, digital photo cameras, webcams, or even cell phone cameras.

The wide availability of low cost printing and reading devices makes 2-D bar codes a very cheap and attractive technology for various multimedia management applications. For example, 2-D bar codes are being used in new emerging applications such as M-ticketing, where they carry selected information of a ticket that is received via a mobile phone [8]; digital postage for online postal services [9]; and automatic tracing and tracking of printed documents such as bills, reports, tax forms, etc. Furthermore, since the halftoning printing technique unavoidably leads to a loss in quality of printed documents, 2-D bar codes are being considered as auxiliary channel dedicated to convey additional side information for improving the quality of scanned documents [10].

Another important field where 2-D bar codes are used is multimedia security. In multimedia security applications, 2-D bar codes are mainly considered as auxiliary data storage means carrying special information for reliable person identification [11] and document authentication systems.

In reliable person identification systems based on identification documents such as passports, identification documents (IDs), driver licences, etc., the personal information (e.g. name, date and place of birth, nationality), the document’s identifier (e.g. ID number), and the personal biometrics (e.g. fingerprint, 2-D photo, three-dimensional scan of the head, iris, sample of the voice or other person-dependent data) are properly encoded and stored into a 2-D bar code. The purpose of the 2-D bar code is twofold. Firstly, it makes the process of information reading from the identification document fast and robust to possible distortions. Secondly, it provides an additional level of security since the
data can be stored in encrypted form using an appropriate cipher system. Thus, the data stored into the 2-D bar code serves as an additional link that connects a person with its physical identification document. Obviously, the storage rate of the 2-D bar code should be adequate to satisfy the storage requirements of both the personal biometrics and the cipher system. Taking into account the growing number of biometrics as well as the particularities of common off-the-shelf cipher systems, the data storage rate of the 2-D bar code should be very high. Alternatively, specific techniques of source coding and feature extraction should be applied to the raw biometric data as well as make use of recent results in elliptic curve cryptography in order to reduce the payload to be stored into the 2-D bar code [12]. In many applications, the 2-D bar code is also required to be inseparable or adherent to the document’s surface in order to reinforce the link between the physical document and its owner as well as to avoid document duplication. Laser engraving of the 2-D bar code seems to be a reasonable solution to this problem.

In document authentication, 2-D bar codes can be used to store either the document’s content or its hashed value. In fact, they constitute a very attractive solution in many document authentication applications where the storage rate of digital data hiding techniques is too low due to the strict constraints imposed on the document’s allowable distortion, or the particularities of some document reproduction and acquisition devices. Moreover, in those applications where the esthetical requirements are of great importance, 2-D bar codes can be printed using special invisible inks or crystals that can be then excited by infrared or ultraviolet light and scanned in the visible range.

Being cheap, simple and possibly exhibiting good security properties, current 2-D bar codes do not offer enough data storage rate for these new applications. In part, this is because most of them use only black and white (B&W) 2-D symbols for encoding data and corresponding binary coding technology. Examples of such 2-D bar codes are Data Matrix, PDF417, Datastrip Code, QR code, etc [13]. Only few proposals exist (commercial and non-commercial) that use multiple gray levels or colors for the 2-D symbols. We call this type of symbologies multilevel 2-D bar codes. Although multilevel 2-D bar codes can potentially increase the achievable rates, in bytes per square inch (bytes/in²), of B&W symbologies, little research has been done on how to efficiently design and implement this approach. Therefore, the main goal of this Chapter is to give a number of guidelines for the design of cheap high-rate multilevel 2-D bar codes.

This Chapter is organized as follows. The description of the problem framework and an evaluation of the performance limits of multilevel 2-D bar codes is given in Section 2.2. A new discrete model for the print-scan channel, specifically adapted to the problem of communications using multilevel 2-D bar codes, is presented in Section 2.3. A review of multilevel coding with multistage decoding for the additive white Gaussian noise channel and the adaptation of this coded modulation technique to the print-scan channel is presented in Section 2.4. Experimental results on the design of practical multilevel 2-D bar codes are presented in Section 2.5. Finally, Section 2.6 concludes the Chapter.
2.2 Framework and Theoretical Performance Limits of Multilevel 2-D Bar Codes

The goal of this Section is to evaluate the theoretical performance limits of multilevel 2-D bar codes in terms of the maximum number of levels that can be placed within a 2-D symbol. Here and hereafter the print-scan (P-S) channel is studied only for the case of B&W halftone printers and low-resolution CCD-based scanners (up to 600 ppi by taking into account the limitations of existing equipment at places such as airport border controls). Therefore, all our results apply only to this specific case. However, our approach can be readily extended and applied to other P-S channels, including those that use color printing and laser engraving as printing technologies, and those that use high-resolution scanners or handheld digital cameras.

Halftone printers simulate multiple gray levels by using the so-called halftoning technique [14]. Assume that both an ideal halftone printer and an ideal scanner are exploited in order to avoid the interference between adjacent 2-D symbols. Furthermore, suppose perfect synchronization, meaning that all the 2-D symbols are accurately read from their exact location on the paper. Let \( r_p \) represent the printer’s resolution measured in dots per inch (dpi), \( a \) be the length in dots of the side of a square halftone cell\(^1\) (see Figure 2.1a), and \( r_s \) denote the scanner’s resolution measured in pixels per inch (ppi). In this case, the printer produces \((r_p/a)^2\) halftone cells per square inch (see Figure 2.1b), each of them capable of representing up to \(a^2 + 1\) gray levels [14, p. 5]. Assuming we use one halftone cell to represent one multilevel 2-D symbol, we can place up to:

\[
U = (r_p/a)^2 \cdot \log_2(a^2 + 1)
\]  

(2.1)

information bits per square inch.

Notice that if \( r_p \) is fixed, \( U \) is a strictly decreasing function of \( a \) (see Figure 2.2) since \( a^2 \) increases faster than \( \log_2(a^2 + 1) \). At first, this fact may appear contradictory since one expects an increase in the storage rate of a multilevel 2-D bar code if the number of employed gray levels is higher. However, we observe from (2.1) that, due to the printing technology, the gain in terms of storage rate of using smaller 2-D symbols is greater than the gain obtained by using multiple gray levels.

\(^1\)The quotient \( r_p/a \) (see Figure 2.1b) is usually known as the printer’s screen frequency and is measured in lines per inch (lpi).
Figure 2.2: Maximum number $U$ of bits per square inch for a multilevel 2-D bar code in the ideal setup.

The minimum scanner resolution to read a 2-D symbol in this ideal case is $r_s = r_p/a$. Indeed, in this case, each scanned pixel corresponds to a 2-D symbol. In practice, however, a minimum resolution of $r_s = k(r_p/a)$, with $k > 1$, is needed in order to obtain both good synchronization and mitigate the inter-symbol interference (ISI). In this scenario, each 2-D symbol is represented by a number of pixels and only the outermost pixels of a 2-D symbol suffer from ISI. Therefore, even if (2.1) is maximized for $a = 1$, which corresponds to a bilevel (BL) 2-D bar code, the minimum scanner resolution $r_s^{BL}$ required for this case may be prohibitively high. On the other hand, if $a > 1$, which corresponds to a multilevel (ML) 2-D bar code using $M = a^2 + 1$ gray levels, then the minimum scanner resolution $r_s^{ML}$ required for this case is $a$ times smaller than $r_s^{BL}$. This fact constitutes a major advantage of multilevel 2-D bar codes over their bilevel counterparts since for the former ones the scanner resolution can be low enabling the use of cheap scanners or the digital cameras present in mobile devices.

### 2.3 Discrete Print-Scan Channel Model

We consider the problem of data transmission via the continuous P-S channel as a digital communications problem.

The continuous P-S channel is such that it introduces several types of distortions, specifically, luminance transformations, scaling, rotation, low pass filtering, aliasing, and noise. Furthermore, its behavior depends on the selected image resolution $r_{im}$ (in ppi); the
parameters used for printing, namely resolution \( r_p \) (in dpi), screen frequency (in lpi), and halftoning algorithm; and the parameters used for scanning, namely resolution \( r_s \) (in ppi), bit-depth, tone correction (e.g. \( \gamma \)-correction), and driver filtering (e.g. descreening filter, that generally acts as a low-pass filter).

We create multilevel 2-D bar codes as digital images. Without loss of generality, we only consider the case in which \( r_s \geq r_{im} \), i.e. the obtained image after printing and scanning has at least the same number of pixels as the original image\(^2\). Although a discrete model for the continuous P-S channel has been proposed for the case where \( r_s = r_{im} \) [15, 16], we do not use it here for two reasons. Firstly, the parameters in this model are difficult to evaluate. Secondly, we are interested in a model including the effects of both a modulator and a demodulator of multilevel 2-D symbols. The discrete channel model we propose assumes that \( r_s > r_{im} \). It models the effects of a modulator, the continuous P-S channel and a demodulator. It is, however, a simplified model since it does not take into account two impairments of the continuous P-S channel, namely desynchronization and interference between 2-D symbols. Still, our discrete channel model shares similar properties with the one proposed in [15, 16], e.g. the dependence between the channel noise and the channel input.

A block diagram showing the different elements considered in our model is depicted in Figure 2.3. Firstly, given a gray value \( X \), the modulator creates a 2-D symbol (e.g. a \( 2 \times 2 \) pixel square) with gray value \( X \). Then, the 2-D symbol passes through the continuous P-S channel. Lastly, the demodulator calculates the average gray value \( Y \) of the received 2-D symbol (e.g. a \( 4 \times 4 \) pixel square if \( r_s = 2 \cdot r_{im} \)).

In the rest of this Chapter and without loss of generality, we fix the shape of a 2-D symbol to be a square. Moreover, we identify a 2-D symbol with a pulse. Therefore, the modulation technique described above is identified with pulse amplitude modulation (PAM) [16].

### 2.3.1 Characterization of the Discrete Print-Scan Channel

In order to reveal the behavior of the discrete P-S channel for multilevel 2-D bar codes, a series of experiments have been performed. We exploited 5 printers and 3 scanners from various manufacturers (see Table 2.1). All the printers were used at their highest

\[ \text{Notice that the printer’s resolution parameter } r_p \text{ determines only the quality of printed reproduction of gray levels whereas the image resolution parameter } r_{im} \text{ determines the size of the printed 2-D symbols.} \]
resolution, default screen frequency, and default halftoning algorithm. All the scanners were used at a resolution of \( r_s = 600 \) ppi, in grayscale mode, and bit-depth set to 8 bits. All the other scanning parameters were set to their default values, however, automatic tone correction and filtering were switched off.

Furthermore, we used \( 2 \times 2 \) pixel square 2-D symbols and 1 pixel of inter-symbol space in order to avoid ISI. The image resolution parameter of all our digital images was set to \( r_{im} = 200 \) ppi. In the scope of this work we did not consider any synchronization algorithm\(^3\). However, in order to deal with rotation and scaling, two typical impairments of the continuous P-S channel, we manually rotated the scanned images and subsequently performed a bicubic interpolation so as to obtain the desired size for these images. As an example, we show in Figure 2.4 the original and noisy versions of a multilevel 2-D bar code (laser) printed-and-scanned under the above conditions.

![Figure 2.4: Multilevel 2-D bar code. (a) Original digital image: \( 2 \times 2 \) pixel symbols, 1 pixel of inter-symbol space, \( r_{im} = 200 \) ppi. (b) Printed-and-scanned digital image: \( r_p = 600 \) dpi, \( r_s = 600 \) ppi, \( 6 \times 6 \) pixel noisy symbols.](image_url)

For the characterization of the discrete P-S channel, all the gray levels from 0 (black) to 255 (white) were used. For each gray level \( x \in [0, 255] \), \( J = 1024 \) 2-D symbols with gray level \( x \) were sent over the continuous P-S channel. Then, the sample mean \( \hat{\mu}_{Y|X}(x) \) and sample variance \( \hat{\sigma}_{Y|X}^2(x) \) of the received noisy symbols \( y_j(x) \in \mathbb{R}, j = 1, \ldots, J, \)

\[^3\text{Nevertheless, this issue was dealt with in the Master thesis of Sarikaya [17], which was cosupervised by the author. The implemented synchronization algorithm being quite trivial, we do not report it here.}\]

<table>
<thead>
<tr>
<th>Model</th>
<th>Type</th>
<th>Code</th>
</tr>
</thead>
<tbody>
<tr>
<td>HP Color LaserJet 4600 (B&amp;W mode)</td>
<td>laser printer</td>
<td>( p_1 )</td>
</tr>
<tr>
<td>HP LaserJet 4350</td>
<td>laser printer</td>
<td>( p_2 )</td>
</tr>
<tr>
<td>Lexmark C760 (color mode)</td>
<td>laser printer</td>
<td>( p_3 )</td>
</tr>
<tr>
<td>OKI B63000</td>
<td>laser printer</td>
<td>( p_4 )</td>
</tr>
<tr>
<td>HP Color DeskJet 990Cxi (color mode)</td>
<td>inkjet printer</td>
<td>( p_5 )</td>
</tr>
<tr>
<td>Epson Perfection 3170 Photo</td>
<td>CCD scanner</td>
<td>( s_1 )</td>
</tr>
<tr>
<td>Canon LiDE 50</td>
<td>CCD scanner</td>
<td>( s_2 )</td>
</tr>
<tr>
<td>HP ScanJet 5300C</td>
<td>CCD scanner</td>
<td>( s_3 )</td>
</tr>
</tbody>
</table>

Table 2.1: Printers and scanners used for experimentation.
Figure 2.5: Sample mean $\hat{\mu}_{Y|X}(x)$ for different combinations of printers and scanners (see Table 2.1): (a) $p_1 - s_1$ (b) $p_1 - s_2$ (c) $p_1 - s_3$ (d) $p_2 - s_1$ (e) $p_2 - s_2$ (f) $p_2 - s_3$ (g) $p_3 - s_1$ (h) $p_3 - s_2$ (i) $p_3 - s_3$ (j) $p_4 - s_1$ (k) $p_4 - s_2$ (l) $p_4 - s_3$ (m) $p_5 - s_1$ (n) $p_5 - s_2$ (o) $p_5 - s_3$. 
Figure 2.6: Sample standard deviation $\hat{\sigma}_{Y|X}(x)$ for different combinations of printers and scanners (see Table 2.1): (a) $p_1 - s_1$ (b) $p_1 - s_2$ (c) $p_1 - s_3$ (d) $p_2 - s_1$ (e) $p_2 - s_2$ (f) $p_2 - s_3$ (g) $p_3 - s_1$ (h) $p_3 - s_2$ (i) $p_3 - s_3$ (j) $p_4 - s_1$ (k) $p_4 - s_2$ (l) $p_4 - s_3$ (m) $p_5 - s_1$ (n) $p_5 - s_2$ (o) $p_5 - s_3$. 
2.3 Discrete Print-Scan Channel Model

were computed\(^4\). The demodulation algorithm (see Figure 2.3) that was used consisted in averaging the gray values of all but the borderline pixels of a noisy 2-D symbol. We did not take into account the borderline pixels in order to reduce desynchronization problems and to avoid ISI. The choice \( r_s = 600 \) ppi is also justified by the same reasons.

We show in Figures 2.5 and 2.6 the obtained results for each printer and scanner combination of Table 2.1. From Figure 2.5, we observe that the typical sample mean function is not linear but can be assumed to be linear in the middle range of gray values. Moreover, it shows saturation effects especially in the lower range (near black) of gray values. From Figure 2.6, we observe that for the tested laser printers \((p_1 \text{ to } p_4)\) the typical sample standard deviation varies according to the employed gray level. However, we also notice that for the tested inkjet printer \((p_5)\) the sample standard deviation tends to be more or less constant for most of the employed gray levels.

### 2.3.2 Characterization of the Noise Probability Distribution

In this Section, we study the problem of characterization of the noise probability distribution. We consider signal constellations composed of eight gray levels (see Table 2.2). The procedure used to select these signal constellations is explained in Sect. 2.5.1.

<table>
<thead>
<tr>
<th>(\mathcal{X})</th>
<th>(p_1 - s_1)</th>
<th>(p_1 - s_2)</th>
<th>(p_1 - s_3)</th>
<th>(p_2 - s_1)</th>
<th>(p_2 - s_2)</th>
<th>(p_2 - s_3)</th>
<th>(p_3 - s_1)</th>
<th>(p_3 - s_2)</th>
<th>(p_3 - s_3)</th>
<th>(p_4 - s_1)</th>
<th>(p_4 - s_2)</th>
<th>(p_4 - s_3)</th>
<th>(p_5 - s_1)</th>
<th>(p_5 - s_2)</th>
<th>(p_5 - s_3)</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>87</td>
<td>116</td>
<td>138</td>
<td>166</td>
<td>196</td>
<td>224</td>
<td>255</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>0</td>
<td>88</td>
<td>118</td>
<td>138</td>
<td>161</td>
<td>187</td>
<td>213</td>
<td>255</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>0</td>
<td>80</td>
<td>113</td>
<td>134</td>
<td>155</td>
<td>182</td>
<td>211</td>
<td>255</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>0</td>
<td>54</td>
<td>88</td>
<td>119</td>
<td>153</td>
<td>190</td>
<td>231</td>
<td>255</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>0</td>
<td>59</td>
<td>96</td>
<td>131</td>
<td>161</td>
<td>199</td>
<td>234</td>
<td>255</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>0</td>
<td>59</td>
<td>95</td>
<td>129</td>
<td>160</td>
<td>197</td>
<td>234</td>
<td>255</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>0</td>
<td>15</td>
<td>35</td>
<td>65</td>
<td>113</td>
<td>174</td>
<td>230</td>
<td>255</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>0</td>
<td>10</td>
<td>27</td>
<td>50</td>
<td>94</td>
<td>161</td>
<td>225</td>
<td>255</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>0</td>
<td>14</td>
<td>33</td>
<td>62</td>
<td>110</td>
<td>174</td>
<td>231</td>
<td>255</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>0</td>
<td>54</td>
<td>93</td>
<td>136</td>
<td>181</td>
<td>218</td>
<td>241</td>
<td>255</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>0</td>
<td>63</td>
<td>105</td>
<td>154</td>
<td>190</td>
<td>222</td>
<td>243</td>
<td>255</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>0</td>
<td>52</td>
<td>90</td>
<td>131</td>
<td>181</td>
<td>217</td>
<td>241</td>
<td>255</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>0</td>
<td>45</td>
<td>94</td>
<td>137</td>
<td>173</td>
<td>204</td>
<td>230</td>
<td>255</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>0</td>
<td>36</td>
<td>76</td>
<td>123</td>
<td>164</td>
<td>197</td>
<td>227</td>
<td>255</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>0</td>
<td>31</td>
<td>67</td>
<td>114</td>
<td>156</td>
<td>193</td>
<td>225</td>
<td>255</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Table 2.2: Tested signal constellations for characterizing the noise probability distribution.

We assigned uniform a-priori probabilities to the signal constellation points and sent the corresponding 2-D bar code symbols \( J = 8192 \) times over the continuous P-S channel. We show in Figure 2.7 the normalized histograms of the received signal points for all printer and scanner combinations. We observe from this figure that given a signal point

\(^4\)The following maximum likelihood estimators were used: \( \hat{\mu}_{Y|X}(x) = \frac{1}{J} \sum_{j=1}^{J} y_j(x) \) and \( \hat{\sigma}_{Y|X}^2(x) = \frac{1}{J} \sum_{j=1}^{J} (y_j(x) - \hat{\mu}_{Y|X}(x))^2. \)
Figure 2.7: Normalized histogram of the received signal points and GGD approximation for each $x \in \mathcal{X}$ (see Tables 2.2 and 2.3): (a) $p_1 - s_1$ (b) $p_1 - s_2$ (c) $p_1 - s_3$ (d) $p_2 - s_1$ (e) $p_2 - s_2$ (f) $p_2 - s_3$ (g) $p_3 - s_1$ (h) $p_3 - s_2$ (i) $p_3 - s_3$ (j) $p_4 - s_1$ (k) $p_4 - s_2$ (l) $p_4 - s_3$ (m) $p_5 - s_1$ (n) $p_5 - s_2$ (o) $p_5 - s_3$. 
2.3 Discrete Print-Scan Channel Model

\( x \in \mathcal{X} \), the conditional probability distribution \( f_{Y|X}(\cdot|x) \) of \((Y|X = x)\) may be modeled, as first approximation, by a \textit{generalized Gaussian distribution} (GGD) [18] with mean \( \mu_{Y|X}(x) \), standard deviation \( \sigma_{Y|X}(x) \), and shape parameter \( \gamma_{Y|X}(x) \), i.e. \((Y|X = x) \sim \mathcal{GGD}(\mu_{Y|X}(x), \sigma_{Y|X}(x), \gamma_{Y|X}(x)) \). The corresponding estimated parameters are shown in Table 2.3.

As we will see in Section 2.5, the main justification of the proposed model is that it gives good results for the problem of data transmission over the continuous P-S channel via multilevel 2-D bar codes.

2.3.3 Discrete Print-Scan Channel Model

Based on our experimental results, reported in Sections 2.3.1 and 2.3.2, we model the P-S channel as (see also Figure 2.8):

\[
Y = \rho(X) + Z, \quad X \in \mathcal{X} = [0, 255], \quad Y \in \mathcal{Y} = \mathbb{R}, \quad (2.2)
\]

where \( X \) (the channel input) represents the gray value of a 2-D symbol, \( \rho : \mathcal{X} \rightarrow \mathbb{R} \) is a nonlinear function representing the response of the P-S channel, \( Z \) represents zero-mean additive noise, and \( Y \) (the channel output) represents the obtained gray value of the corresponding 2-D symbol. Gray values are represented as numbers in the set \( \{0, 1, \ldots, 255\} \).

The function \( \rho(\cdot) \) is in general different for every particular instance of the P-S channel, i.e. for every printer and scanner combination. Contrary to what is usually assumed, the noise term \( Z \) is not supposed to be independent of the gray value \( X \). This is in agreement with the discrete model proposed in [15, 16]. In fact, according to the experimental results reported in Sect. 2.3.2, we suppose that for each channel use \( Z \) is drawn i.i.d. from a zero-mean generalized Gaussian distribution with standard deviation and shape parameter depending on the channel input \( X \). Thus, we model \((Z|X = x)\) as:

\[
(Z|X = x) \sim \mathcal{GGD}(0, \sigma_{Z|X}(x), \gamma_{Z|X}(x)) \quad (2.3)
\]

\( \text{Recall that if } X \sim \mathcal{GGD}(\mu, \sigma, \gamma), \text{ then the p.d.f. of } X \text{ is given by:} \\ f_X(x) = \frac{\gamma \eta(\sigma, \gamma)}{2 \Gamma(1/\gamma)} \exp\left(-\frac{1}{\gamma} \frac{1}{\gamma} \left| \frac{x - \mu}{\gamma} \right| \right), \\ \text{where } \mu, \sigma, \text{ and } \gamma \text{ denote, respectively, the mean, the standard deviation, and the shape parameter of } X, \\ \Gamma(\cdot) \text{ denotes the Gamma function, and:} \\ \eta(\sigma, \gamma) = \frac{1}{\sigma} \sqrt{\frac{\Gamma(3/\gamma)}{\Gamma(1/\gamma)}}. \)
| x   | $\mu_{Y|X}(x)$ | $\sigma_{Y|X}(x)$ | $\gamma_{Y|X}(x)$ | x   | $\mu_{Y|X}(x)$ | $\sigma_{Y|X}(x)$ | $\gamma_{Y|X}(x)$ | x   | $\mu_{Y|X}(x)$ | $\sigma_{Y|X}(x)$ | $\gamma_{Y|X}(x)$ |
|-----|----------------|------------------|------------------|-----|----------------|------------------|------------------|-----|----------------|------------------|------------------|
| 0   | 57.9           | 4.20             | 2.07             | 0   | 97.56          | 11.01            | 1.79             | 0   | 44.75          | 27.73            | 1.51             |
| 87  | 88.7           | 7.48             | 1.64             | 87  | 123.96         | 11.40            | 1.79             | 87  | 91.88          | 21.52            | 1.63             |
| 116 | 122.1          | 8.43             | 1.68             | 116 | 150.12         | 10.45            | 2.04             | 116 | 131.43         | 16.50            | 1.70             |
| 138 | 155.8          | 8.98             | 1.98             | 138 | 179.13         | 9.62             | 2.31             | 138 | 165.34         | 13.60            | 2.08             |
| 166 | 182.6          | 7.62             | 1.63             | 166 | 200.57         | 7.51             | 2.12             | 166 | 191.31         | 10.11            | 2.09             |
| 196 | 209.3          | 5.69             | 2.00             | 196 | 217.50         | 5.75             | 1.99             | 196 | 214.77         | 6.55             | 2.05             |
| 224 | 229.4          | 4.15             | 2.06             | 224 | 230.55         | 4.39             | 1.97             | 224 | 229.49         | 4.51             | 1.90             |
| 255 | 241.3          | 1.75             | 1.92             | 255 | 240.88         | 3.80             | 1.83             | 255 | 240.69         | 3.59             | 1.84             |

(a) | (b) | (c) | (d) | (e) | (f) | (g) | (h) | (i) | (j) | (k) | (l) | (m) | (n) | (o) |
Table 2.3: Estimated parameters for the GGD approximation of the noise probability distribution: (a) $p_1 - s_1$ (b) $p_1 - s_2$ (c) $p_1 - s_3$ (d) $p_2 - s_1$ (e) $p_2 - s_2$ (f) $p_2 - s_3$ (g) $p_3 - s_1$ (h) $p_3 - s_2$ (i) $p_3 - s_3$ (j) $p_4 - s_1$ (k) $p_4 - s_2$ (l) $p_4 - s_3$ (m) $p_5 - s_1$ (n) $p_5 - s_2$ (o) $p_5 - s_3$. 
where, \(\sigma_{Z|X}(x)\) and \(\gamma_{Z|X}(x)\) are, respectively, the noise standard deviation and noise shape parameter given that \(X = x\). As a final remark, notice that our discrete channel model assumes perfect synchronization and absence of ISI.

For practical schemes based on this model, the channel response \(\rho(\cdot)\) is approximated by the sample mean \(\hat{\mu}_{Y|X}(\cdot)\), i.e. \(\rho(x) = \hat{\mu}_{Y|X}(x)\), the noise standard deviation \(\sigma_{X|X}(\cdot)\) by the sample standard deviation \(\hat{\sigma}_{Y|X}(\cdot)\), i.e. \(\sigma_{Z|X}(x) = \hat{\sigma}_{Y|X}(x)\), and the noise shape parameter \(\gamma_{Z}(\cdot)\) by the sample shape parameter \(\hat{\gamma}_{Y|X}(\cdot)\), i.e. \(\gamma_{Z|X}(x) = \hat{\gamma}_{Y|X}(x)\).

## 2.4 Multilevel Coding for the Print-Scan Channel

In Sect. 2.3.3, we modeled the P-S channel as a (discrete) channel with luminance-dependent generalized Gaussian additive noise. Moreover, we showed in Sect. 2.3.2 that the estimated noise shape parameters are close to 2, which corresponds to a Gaussian distribution of the noise. In this Section, inspired by the results obtained in [19] for the additive white Gaussian noise (AWGN) channel, we study the suitability of multilevel coding (MLC) together with multistage decoding (MSD) for the P-S channel. Thus, the goal of this Section is twofold. Firstly, we consider the fundamentals of MLC/MSD for the discrete AWGN channel. Secondly, we adapt the theory of multilevel coding to the case of the discrete P-S channel modeled by (2.2) and (2.3).

### 2.4.1 MLC/MSD for the AWGN Channel

The idea of coded modulation is to jointly optimize coding and modulation in order to improve the performance of digital communication schemes [20]. MLC/MSD is a well-known bandwidth-efficient scheme for the AWGN channel in the high signal-to-noise ratio (SNR) regime [21]. In the following paragraphs we recall the fundamental results concerning MLC/MSD.

Consider the discrete AWGN channel model:

\[
Y = X + Z, \quad X \in \mathcal{X}, \quad Y \in \mathcal{Y}, \quad Z \sim \mathcal{N}(0, \sigma_Z^2),
\]

where \(X\) (respectively \(Y\)) is the channel input (respectively output), \(\mathcal{X}\) (respectively \(\mathcal{Y}\)) is the input (respectively output) alphabet, and \(Z\) represents the noise. For each channel use, the noise \(Z\) is drawn i.i.d. from a zero-mean Gaussian distribution with variance \(\sigma_Z^2\) and is assumed to be independent from the channel input \(X\). This model can be considered as a particular case of model (2.2) and (2.3) under the assumption that \(\rho(x) = x, \sigma_{Z|X}(x) = \sigma_Z\), and \(\gamma_{Z|X}(x) = 2\), for all \(x \in \mathcal{X}\). Given the channel input variance \(\sigma_X^2\), the capacity of this channel, in bits per channel use, is:

\[
C_{AWGN} = \max_{f_X(\cdot)} I(X; Y) = \frac{1}{2} \log_2 \left(1 + \frac{\sigma_X^2}{\sigma_Z^2}\right),
\]

which is attained when the channel input \(X\) is Gaussian, for example \(X \sim \mathcal{N}(\mu_X, \sigma_X^2)\).

Due to the technical impossibility of using either a continuous or an infinite input alphabet, practical systems usually employ a discrete and finite \(M = 2^L\)–ary input alphabet (signal constellation), i.e. \(|\mathcal{X}| = 2^L\). For example, we show in Figure 2.9 a PAM signal constellation with \(M = 8\) (\(L = 3\)) equidistant signal points.
Figure 2.9: Transmission of an 8–PAM equidistant constellation over the AWGN channel.

It is customary to assign a binary label to each signal point by means of a bijective mapping $\psi$:

$$\begin{align*}
(b_0^i, b_1^i, \ldots, b_{L-1}^i) &\mapsto x, \\
&\quad \psi
\end{align*}$$

(bi, 0, 1, ..., L − 1)

Given a specific probability distribution $\{p(x) : x \in \mathcal{X}\}$ over the channel inputs, the maximum rate of reliable communications of such systems is given by the mutual information $I(X; Y)$.

Remarkably, MLC/MSD [19, 22, 23] is a straightforward consequence of the chain rule for mutual information. Since the mapping $\psi$ in (2.4) is bijective, the mutual information $I(X; Y)$ between the transmitted signal $X$ and the received signal $Y$ equals the mutual information $I(B_0^i, B_1^i, \ldots, B_{L-1}^i; Y)$ between the binary label of $X$ and the received signal $Y$. Applying the chain rule for mutual information, we get:

$$I(X; Y) = I(B_0^i, B_1^i, \ldots, B_{L-1}^i; Y)$$

$$= I(B_0^i; Y) + I(B_1^i; Y|B_0^i) + \ldots + I(B_{L-1}^i; Y|B_0^i, B_1^i, \ldots, B_{L-2}^i).$$

Equation (2.5) can be interpreted as follows. Transmission of vectors with binary digits $b^i, i = 0, 1, \ldots, L − 1$, over the physical channel can be separated into the parallel transmission of individual bits $b^i$ over $L$ equivalent channels, provided that $b_0^i, b_1^i, \ldots, b_{L-1}^i$ are known (see Figures 2.10 and 2.11).

At the transmitter side (see Figure 2.12), a binary data block of length $K$ bits is partitioned into $L$ sub-blocks:

$$q = (q_1, \ldots, q_K), \quad q_k \in \mathcal{B}, \quad k = 1, \ldots, K,$$

$$q = (q_0^i, \ldots, q_{L-1}^i), \quad q_i^i = (q_1^i, \ldots, q_K^i),$$

$$i = 0, 1, \ldots, L − 1, \quad \sum_{i=0}^{L-1} K_i = K.$$

Each data sub-block $q_i$ is fed into an individual binary encoder $E_i$ of rate $R_i = K_i/N$ producing a codeword:

$$b_i = (b_1^i, \ldots, b_{N}^i), \quad b_n^i \in \mathcal{B},$$

$$n = 1, \ldots, N, \quad i = 0, 1, \ldots, L − 1,$$
2.4 Multilevel Coding for the Print-Scan Channel

Figure 2.10: Conditional p.d.f.’s for the equivalent channel at level 0 when an 8–PAM equidistant constellation with Ungerböck’s labeling is used: (a) $f(y|b^0 = 0)$, (b) $f(y|b^0 = 1)$.

Figure 2.11: Conditional p.d.f.’s for the equivalent channel at level 1 when an 8–PAM equidistant constellation with Ungerböck’s labeling is used: (a) $f(y|b^1 = 0, b^0 = 0)$, (b) $f(y|b^1 = 0, b^0 = 1)$, (c) $f(y|b^1 = 1, b^0 = 0)$, (d) $f(y|b^1 = 1, b^0 = 1)$. 
of the corresponding component code. In this manner $L$ levels of coding are created. For simplicity, we assume that all codewords have equal length, $N$ binary symbols, at all levels. Then, the $n$-th bit $b^n_i$, $n = 1, \ldots, N$, of every codeword $b^i$ is selected to form a binary label $(b^n_0, b^n_1, \ldots, b^{L-1}_n)$ of $L$ bits, which is mapped via $\psi$ to a signal point $x_n \in \mathcal{X}$. In this way, we obtain a vector:

$$x = (x_1, \ldots, x_N), \quad x_n \in \mathcal{X}, \quad n = 1, \ldots, N,$$

of $N$ channel inputs, which are serially transmitted over the AWGN channel. It is very easy to show that the code rate of the overall scheme, $R = K/N$, is equal to the sum of the individual code rates:

$$\sum_{i=0}^{L-1} R_i = \sum_{i=0}^{L-1} \frac{K_i}{N} = \frac{1}{N} \sum_{i=0}^{L-1} K_i = \frac{K}{N} = R.$$

At the receiver side, the component codes are successively decoded by the corresponding decoders starting from the lowest level. At any stage $i$, $i = 0, 1, \ldots, L-1$, the decoder processes not only the $N$ received signal points:

$$y = (y_1, \ldots, y_N), \quad y_n \in \mathcal{Y}, \quad n = 1, \ldots, N,$$
but also decisions of previous decoding stages:

\[ \hat{b}^j = (\hat{b}^j_1, \ldots, \hat{b}^j_N), \quad \hat{b}^j_n \in B, \]

\[ n = 1, \ldots, N, \quad j = 0, 1, \ldots, i - 1. \]

The block diagram of the receiver is shown in Figure 2.13. For simplicity, this diagram represents neither the “Data selection” block, which outputs \( \hat{q}^0, \ldots, \hat{q}^{L-1} \), nor the “Concatenation of data” block, which outputs \( \hat{q} \).

One can demonstrate that the maximum achievable rate of a modulation scheme (e.g. 8–PAM) with given a-priori probabilities of its signal constellation points can indeed be achieved by MLC/MSD if, and only if, the individual rates \( R_i \) of the component codes are chosen to be equal to the capacities of the equivalent channels, i.e:

\[ R_i = I(B^i; Y|B^0, B^1, \ldots, B^{i-1}), \quad i = 0, 1, \ldots, L - 1. \]

This is the so-called capacity rule for choosing the individual code rates [19] (see Figure 2.14).

![Figure 2.14: Capacity rule for selecting the individual code rates.](image)

The capacity rule does not indicate however which codes to use for the equivalent channels. In practice, since the statistics of the equivalent channels \( f(y|b^i, b^0, \ldots, b^{i-1}) \) are highly non Gaussian (see Figures 2.10 and 2.11) and no optimal codes are known for these channels, one has some freedom for selecting the individual codes. In previous
work, it was shown that very long Turbo codes and low density parity-check (LDPC) codes perform very well on MLC/MSD schemes for the AWGN channel [24, 25].

Moreover, since the capacity rule applies to any labeling, there is no restriction on the particular labeling used in MLC/MSD. Nevertheless, for finite codeword length, Ungerbök’s labeling turns out to show the highest performance among MLC/MSD schemes with different labelings [19].

2.4.2 MLC/MSD for the Print-Scan Channel

Suppose we use an $M$-ary ($|\mathcal{X}| = M = 2^L$) modulation system with given a-priori probabilities of the signal constellation points \( \{p(x): x \in \mathcal{X}\} \). Since our channel is memoryless, i.e. the output $Y$ depends solely on the current channel input $X$ and current noise term $Z$, the maximum achievable rate of this modulation system is given by the mutual information $I(X; Y)$. Although we cannot provide a closed-form expression for $I(X; Y)$ it is fairly easy to compute it numerically. By definition:

$$I(X; Y) = h(Y) - h(Y/X). \quad (2.6)$$

Firstly, we can compute the term $h(Y)$ in (2.6) as follows. By definition:

$$h(Y) = -\int f(y) \log f(y) \, dy. \quad (2.7)$$

In order to numerically evaluate the integral in (2.7), we need to specify $f(y)$. This can be done by marginalizing $f(x, y) = f(y|x)p(x)$:

$$f(y) = \sum_{x \in \mathcal{X}} f(y|x)p(x). \quad (2.8)$$

The term $f(y|x)$ in (2.8) can be obtained by noting that $\{Y|X = x\}$ has a generalized Gaussian distribution with mean $\rho(x)$, standard deviation $\sigma_{Z|X}(x)$, and shape parameter $\gamma_{Z|X}(x)$, i.e. $\{Y|X = x\} \sim GGPDF(\rho(x), \sigma_{Z|X}(x), \gamma_{Z|X}(x))$.

Secondly, the term $h(Y/X)$ in (2.6) can be computed as follows:

$$h(Y/X) = h(\rho(X) + Z/X) = h(Z/X) = \sum_{x \in \mathcal{X}} h(Z/X = x)p(x),$$

where we used (2.2) and the definition of conditional differential entropy [26]. Since $\{Z/X = x\}$ has a generalized Gaussian distribution with zero mean, standard deviation $\sigma_{Z|X}(x)$, and shape parameter $\gamma_{Z|X}(x)$, we have [27]:

$$h(Z/X = x) = \log_2 \left( \frac{2 \cdot \Gamma(1/\gamma_{Z|X}(x)) \exp(1/\gamma_{Z|X}(x))}{\gamma_{Z|X}(x) \eta(\sigma_{Z|X}(x), \gamma_{Z|X}(x))} \right).$$

Therefore,

$$h(Y/X) = \sum_{x \in \mathcal{X}} \log_2 \left( \frac{2 \cdot \Gamma(1/\gamma_{Z|X}(x)) \exp(1/\gamma_{Z|X}(x))}{\gamma_{Z|X}(x) \eta(\sigma_{Z|X}(x), \gamma_{Z|X}(x))} \right) p(x)$$

$$= Ep_x \left[ \log_2 \left( \frac{2 \cdot \Gamma(1/\gamma_{Z|X}(X)) \exp(1/\gamma_{Z|X}(X))}{\gamma_{Z|X}(X) \eta(\sigma_{Z|X}(X), \gamma_{Z|X}(X))} \right) \right].$$
Although we are able to compute $I(X; Y)$ for any given input distribution $p_X(\cdot)$, the problem of finding the capacity $\max_{p_X(\cdot)} I(X; Y)$ of the discrete P-S channel is more involved. The difficulty arises from the fact that both terms in the right-hand-side of (2.6) depend on $p_X(\cdot)$.

However, we can follow the same reasoning and notation as in Sect. 2.4.1 and still make use of MLC/MSD in order to approach $I(X; Y)$. The main difference with respect to the discrete AWGN channel is that the noise $Z$ has a generalized Gaussian distribution with standard deviation and shape parameter depending on the channel input $X$. We show in Figures 2.15 and 2.16 an example of the equivalent channels corresponding to a P-S channel modeled by (2.2) and (2.3). For $i = 0, 1, \ldots, L - 1$, the individual rates $R_i$ of

Figure 2.15: Conditional p.d.f.’s for the equivalent channel at level 0 when an 8–PAM non-equidistant constellation with Ungerboeck’s labeling is used for communications over a P-S channel: (a) $f(y|b^0 = 0)$, (b) $f(y|b^0 = 1)$.

the component codes can be computed as follows:

$$R_i = I(B^i; Y|B^0, \ldots, B^{i-1})$$

$$= h(Y|B^0, \ldots, B^{i-1}) - h(Y|B^0, \ldots, B^{i-1}, B^i).$$

(2.9)

We compute the term $h(Y|B^0, \ldots, B^{i-1})$ in (2.9) as follows. By definition:

$$h(Y|B^0, \ldots, B^{i-1}) = \sum_{(b^0, \ldots, b^{i-1}) \in B^i} h(Y|B^0 = b^0, \ldots, B^{i-1} = b^{i-1})p(b^0, \ldots, b^{i-1}),$$

and

$$h(Y|B^0 = b^0, \ldots, B^{i-1} = b^{i-1}) = -\int_Y f(y|b^0, \ldots, b^{i-1}) \log f(y|b^0, \ldots, b^{i-1}) dy.$$ 

Hence, in order to compute $h(Y|B^0 = b^0, \ldots, B^{i-1} = b^{i-1})$ we need to specify

$$f(y|b^0, \ldots, b^{i-1}).$$

This can be done by conditioning on $B^i = b^i, \ldots, B^{L-1} = b^{L-1}$:

$$f(y|b^0, \ldots, b^{i-1}) = \sum_{(b^0, \ldots, b^{L-1}) \in B^{L-i}} f(y|b^0, \ldots, b^{i-1}, b^i, \ldots, b^{L-1}) \cdot p(b^i, \ldots, b^{L-1}).$$
Therefore,

\[ f(y|b^0, \ldots, b^{i-1}) = \sum_{(b^0, \ldots, b^{L-1}) \in B^{L-1}} f(y|\psi(b^0, \ldots, b^{L-1})) \cdot p(b^i, \ldots, b^{L-1}). \]

Exactly like for \( f(y) \) in (2.8), \( f(y|\psi(b^0, \ldots, b^{L-1})) \) can be obtained noting that

\[ (Y|X = x) \sim \mathcal{G}\mathcal{D}(\rho(x), \sigma_{Z|X}(x), \gamma_{Z|X}(x)). \]

Obviously, we can proceed in the same manner in order to compute the term

\[ h(Y|B^0, \ldots, B^{i-1}, B^i) \]

in (2.9). Notice that since \( f(y|\psi(b^0, \ldots, b^{L-1})) \) depends on the actual mapping \( \psi \), \( R_i \) in (2.9) is also mapping dependent for \( i = 0, 1, \ldots, L - 1 \).
2.5 Experimental Results on the Design of Multilevel 2-D Bar Codes Using MLC/MSD

In this Section, we exploit the discrete channel model developed in Sect. 2.3.3 and the results of Sect. 2.4.2 in order to design practical multilevel 2-D bar codes.

2.5.1 Constellation and MLC/MSD Rate Design for Communications Over the Print-Scan Channel

By using the estimated functions $\rho(\cdot)$ and $\sigma_{Z|X}(\cdot)$, we select a signal constellation $\mathcal{X}$ in such a way that the received signal points are close to each other where the noise standard deviation is small and farther apart where the noise standard deviation is large. The actual number of selected constellation points depends on the first- and second-order statistics of the P-S channel and the error-correcting capabilities of the individual codes used for MLC/MSD. The selected non-equidistant 8–PAM signal constellations for the considered combinations of printers and scanners are reported in Table 2.2.

For illustration, we schematize in Figure 2.17 the employed algorithm for building the constellation for $p_2 - s_3$. If we define $\Delta = \rho(255) - \rho(0)$, then the constellation points and the value of $k$ in this figure can be obtained by solving the following minimization problem:

$$
\min_{x_1, \ldots, x_8} \min_{k \in \mathbb{R}^+} \left| \Delta - k \left( \sigma_{Z|X}(x_1) + 2 \sum_{i=2}^{7} \sigma_{Z|X}(x_i) + \sigma_{Z|X}(x_8) \right) \right|,
$$

subject to $x_1 < x_2 < \ldots < x_8$ and $x_1, \ldots, x_8 \in [0, 255]$.

In Sect. 2.3.2, we explained how to determine the parameters of the discrete P-S channel model for the selected constellation $\mathcal{X}$. The obtained results for the considered printer and scanner combinations are shown in Figure 2.7 and Table 2.3. By using these parameters and following the procedure described in Sect. 2.4.2, we numerically compute the rates $R_i$, $i = 0, 1, 2$, of an MLC/MSD scheme employing the signal constellation $\mathcal{X}$, together with Ungerböck’s labeling. We show in Table 2.4 the obtained rates.

2.5.2 Performance Evaluation of Multilevel 2-D Bar Codes

We studied the performance of multilevel 2-D bar codes using MLC/MSD for the considered printer and scanner combinations.

For this task, we implemented respectively 15 non-equidistant 8–PAM MLC/MSD schemes using the design parameters obtained in Sect. 2.5.1, namely signal constellations (Table 2.2) and channel model parameters (Table 2.3). For the individual component codes of all MLC/MSD schemes, we used LDPC codes. More precisely, we used the Matlab implementation of quasi-regular LDPC codes from I. Kozintsev [28].

A multilevel encoder has a straightforward implementation. For the multistage decoder, one must take into account the statistics of the underlying discrete P-S channel model in order to correctly compute the log-likelihood ratios $l_n^i$ used in the LDPC belief
Figure 2.17: Building up the non-equidistant 8–PAM signal constellation $\mathcal{X}$ for the $p_2 - s_3$ combination, $k = 1.09$.

<table>
<thead>
<tr>
<th></th>
<th>$R_0$</th>
<th>$R_1$</th>
<th>$R_2$</th>
<th>$R$</th>
</tr>
</thead>
<tbody>
<tr>
<td>$p_1 - s_1$</td>
<td>0.8318</td>
<td>0.9993</td>
<td>1.0000</td>
<td>2.8311</td>
</tr>
<tr>
<td>$p_1 - s_2$</td>
<td>0.4153</td>
<td>0.9669</td>
<td>1.0000</td>
<td>2.3822</td>
</tr>
<tr>
<td>$p_1 - s_3$</td>
<td>0.3699</td>
<td>0.9213</td>
<td>0.9996</td>
<td>2.2908</td>
</tr>
<tr>
<td>$p_2 - s_1$</td>
<td>0.2490</td>
<td>0.9158</td>
<td>1.0000</td>
<td>2.1647</td>
</tr>
<tr>
<td>$p_2 - s_2$</td>
<td>0.2083</td>
<td>0.8796</td>
<td>1.0000</td>
<td>2.0880</td>
</tr>
<tr>
<td>$p_2 - s_3$</td>
<td>0.2544</td>
<td>0.9019</td>
<td>1.0000</td>
<td>2.1563</td>
</tr>
<tr>
<td>$p_3 - s_1$</td>
<td>0.6965</td>
<td>0.9947</td>
<td>1.0000</td>
<td>2.6912</td>
</tr>
<tr>
<td>$p_3 - s_2$</td>
<td>0.3244</td>
<td>0.9126</td>
<td>1.0000</td>
<td>2.2370</td>
</tr>
<tr>
<td>$p_3 - s_3$</td>
<td>0.2677</td>
<td>0.7200</td>
<td>0.9915</td>
<td>1.9793</td>
</tr>
<tr>
<td>$p_4 - s_1$</td>
<td>0.1159</td>
<td>0.7060</td>
<td>0.9968</td>
<td>1.8187</td>
</tr>
<tr>
<td>$p_4 - s_2$</td>
<td>0.0827</td>
<td>0.4924</td>
<td>0.9942</td>
<td>1.5693</td>
</tr>
<tr>
<td>$p_4 - s_3$</td>
<td>0.1095</td>
<td>0.6699</td>
<td>0.9952</td>
<td>1.7746</td>
</tr>
<tr>
<td>$p_5 - s_1$</td>
<td>0.9632</td>
<td>0.9999</td>
<td>1.0000</td>
<td>2.9631</td>
</tr>
<tr>
<td>$p_5 - s_2$</td>
<td>0.6286</td>
<td>0.9825</td>
<td>1.0000</td>
<td>2.6110</td>
</tr>
<tr>
<td>$p_5 - s_3$</td>
<td>0.8774</td>
<td>0.9992</td>
<td>1.0000</td>
<td>2.8766</td>
</tr>
</tbody>
</table>

Table 2.4: Individual code rates and total rate for multilevel 2-D bar codes using MLC/MSD ($N \to \infty$).
Experimental Results on the Design of Multilevel 2-D Bar Codes

The propagation decoding algorithm:

\[ l^n_i = \ln \frac{f_{Y|B^i,B^0,...,B^{i-1}}(y_n|1, \hat{b}^0, \ldots, \hat{b}^{i-1})}{f_{Y|B^i,B^0,...,B^{i-1}}(y_n|0, b^0, \ldots, \hat{b}^{i-1})}, \quad i = 0, \ldots, L - 1, \quad n = 1, \ldots, N. \]

The terms \( f_{Y|B^i,B^0,...,B^{i-1}}(y_n|b_i, \hat{b}^0, \ldots, \hat{b}^{i-1}) \), \( b_i = 0, 1 \), in the last equation, are computed as in Sect. 2.4.2.

The bit error rate (BER) of the designed multilevel 2-D bar codes was evaluated for \( R = 2 \) bits per 2-D symbol or, equivalently, 1121 bytes / in\(^2\). The individual code rates \( R_i \), \( i = 0, 1, 2 \), were selected according to the capacity rule for an 8–PAM MLC/MSD scheme with Ungerböck’s labeling over an AWGN channel (see the first row of Table 2.6).

We present the obtained BER results in Table 2.5. For comparison, three channel models, all derived from (2.2) and (2.3), were considered for decoding the multilevel 2-D bar codes. The first model assumes stationary Gaussian noise, i.e., \( Z \sim \mathcal{N}(0, \sigma_Z^2) \), where:

\[ \sigma_Z^2 = \frac{1}{|\mathcal{X}|} \sum_{x \in \mathcal{X}} \sigma_{Z|x}^2(x), \quad \gamma_{Z|x}(x) = 2, \quad \text{for all } x \in \mathcal{X}. \]

The second model assumes luminance-dependent Gaussian noise, meaning that \( \gamma_{Z|x}(x) = 2 \) for all \( x \in \mathcal{X} \), but the noise standard deviation \( \sigma_{Z|x}(x) \) varies for each \( x \in \mathcal{X} \). Finally, the third model assumes luminance-dependent generalized Gaussian noise, meaning that noise standard deviation \( \sigma_{Z|x}(x) \) and shape parameter \( \gamma_{Z|x}(x) \) vary for all \( x \in \mathcal{X} \) as shown in Table 2.3.

### 2.5.3 Fine Tuning of Multilevel 2-D Bar Codes Using MLC/MSD

From the results in Tables 2.4 and 2.5, we observe that the best and worst combinations of laser printer and scanner are, respectively, \( p_1 - s_1 \) and \( p_4 - s_2 \). We notice also that the inkjet printer and scanner combination \( p_5 - s_1 \) is the most attractive since it gives the best results in terms of achievable rate \( R \). In this Section, we investigate the practical performance limits of multilevel 2-D bar codes for the best combinations \( p_1 - s_1 \) and \( p_5 - s_1 \).

We evaluated the performance of the designed multilevel 2-D bar codes for different embedding rates \( R \) (in bits / 2-D symbol). For each embedding rate \( R_i \), the individual code rates \( R_i, \ i = 0, 1, 2 \), were selected by using the capacity rule for an 8–PAM MLC/MSD scheme with Ungerböck’s labeling over an AWGN channel. However, we did not use this selection rule for the embedding rates \( R = 2.8311 \) and \( R = 2.9631 \) which correspond to the rates in the first and thirteenth rows of Table 2.4. For a block length of \( N = 2048 \) bits at all levels, the corresponding generator and parity-check matrices were generated randomly according to these rates. Nevertheless, due to a limitation of the employed software package, we were not able to generate these matrices for very high rates, i.e. for rates \( R_i \) such that \( 0.9951 \leq R_i < 1 \). For these cases, \( R_i = 1 \) was taken. We summarize in Table 2.6 the selected parameters for multilevel coding. Therein, we have \( R_i \approx K_i/N \). The numbers in parentheses correspond to the cases where \( R_i \) was rounded up to 1.

We present the obtained BER results in Tables 2.7a and 2.7b. The last rows in these

---

\(^6\)We do not investigate further the worst combination \( p_4 - s_2 \) because for this case the gain in terms of storage rate of multilevel 2-D bar codes with respect to B&W 2-D bar codes would be minor (see Tables 2.4 and 2.8).
\[
Z \sim \mathcal{N}(0, \sigma_Z^2) \\
(Z|X = x) \sim \mathcal{N}(0, \sigma_Z^2|X(x)) \\
(Z|X = x) \sim \mathcal{GGD}(0, \sigma_Z|X(x), \gamma_Z|X(x))
\]

<table>
<thead>
<tr>
<th></th>
<th>BER mean</th>
<th>BER variance</th>
<th>BER mean</th>
<th>BER variance</th>
<th>BER mean</th>
<th>BER variance</th>
</tr>
</thead>
<tbody>
<tr>
<td>(p_1 - s_1)</td>
<td>0</td>
<td>8.154 \times 10^{-4}</td>
<td>7.641 \times 10^{-6}</td>
<td>9.341 \times 10^{-10}</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>(p_1 - s_2)</td>
<td>2.814 \times 10^{-2}</td>
<td>2.178 \times 10^{-4}</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>(p_1 - s_3)</td>
<td>2.098 \times 10^{-2}</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>(p_2 - s_1)</td>
<td>6.703 \times 10^{-2}</td>
<td>2.487 \times 10^{-4}</td>
<td>1.726 \times 10^{-2}</td>
<td>3.284 \times 10^{-4}</td>
<td>1.680 \times 10^{-2}</td>
<td>3.191 \times 10^{-4}</td>
</tr>
<tr>
<td>(p_2 - s_2)</td>
<td>7.859 \times 10^{-2}</td>
<td>4.429 \times 10^{-4}</td>
<td>1.296 \times 10^{-2}</td>
<td>4.331 \times 10^{-4}</td>
<td>1.052 \times 10^{-2}</td>
<td>2.619 \times 10^{-4}</td>
</tr>
<tr>
<td>(p_2 - s_3)</td>
<td>6.452 \times 10^{-2}</td>
<td>4.429 \times 10^{-4}</td>
<td>1.296 \times 10^{-2}</td>
<td>4.331 \times 10^{-4}</td>
<td>1.052 \times 10^{-2}</td>
<td>2.619 \times 10^{-4}</td>
</tr>
<tr>
<td>(p_3 - s_1)</td>
<td>1.817 \times 10^{-2}</td>
<td>3.821 \times 10^{-4}</td>
<td>4.149 \times 10^{-3}</td>
<td>3.744 \times 10^{-3}</td>
<td>1.032 \times 10^{-5}</td>
<td>0</td>
</tr>
<tr>
<td>(p_3 - s_2)</td>
<td>1.010 \times 10^{-1}</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>(p_3 - s_3)</td>
<td>1.243 \times 10^{-1}</td>
<td>3.432 \times 10^{-5}</td>
<td>1.283 \times 10^{-1}</td>
<td>1.943 \times 10^{-5}</td>
<td>1.276 \times 10^{-1}</td>
<td>1.124 \times 10^{-5}</td>
</tr>
<tr>
<td>(p_4 - s_1)</td>
<td>1.356 \times 10^{-1}</td>
<td>8.754 \times 10^{-5}</td>
<td>1.563 \times 10^{-1}</td>
<td>4.813 \times 10^{-5}</td>
<td>1.649 \times 10^{-1}</td>
<td>2.800 \times 10^{-5}</td>
</tr>
<tr>
<td>(p_4 - s_2)</td>
<td>1.530 \times 10^{-1}</td>
<td>2.478 \times 10^{-5}</td>
<td>1.176 \times 10^{-1}</td>
<td>3.058 \times 10^{-5}</td>
<td>1.168 \times 10^{-1}</td>
<td>4.366 \times 10^{-5}</td>
</tr>
<tr>
<td>(p_4 - s_3)</td>
<td>1.243 \times 10^{-1}</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>(p_5 - s_1)</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>(p_5 - s_2)</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>(p_5 - s_3)</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
</tbody>
</table>

Table 2.5: BER results of multilevel 2-D bar codes for \(R = 2\) (\(N = 2048\)).
Tables correspond to the performance results of the unencoded version of the designed multilevel 2-D bar codes. For this case, maximum likelihood detection was employed along with the above three channel models.

From Table 2.7b, we observe that for the stationary and luminance-dependent Gaussian noise models, the performance of the multistage decoder is quite similar. This is in agreement with Figure 2.6m, where we can see that for the $p_5 - s_1$ combination the noise standard deviation is approximately constant. However, this equivalence in performance is not expected when the noise standard deviation varies (see for example Figure 2.6a and Table 2.3a). Indeed, we observe from the second row of Table 2.7a that for the $p_1 - s_1$ combination the luminance-dependent Gaussian noise model provides a substantial gain in performance with respect to the stationary Gaussian noise model. Tables 2.5, 2.7a and 2.7b also show that the performance of the multistage decoder when using the luminance-dependent generalized Gaussian noise model is in almost all cases better than those obtained for the two other noise models.

Concerning the next to last rows in Tables 2.7a and 2.7b, which correspond to the theoretical maximum achievable rates of the proposed modulation schemes, we think that the main reasons explaining the marginal advantage of the multistage decoder with respect to the unencoded version (last row) are twofold. First, the blocklength $N$ of the underlying codes is simply too small. Second, there is no guarantee that the randomly selected parity-check matrices used for multistage decoding show good performance for the considered individual rates (see Table 2.6).

Notice that the results in Tables 2.7a and 2.7b were obtained for a suboptimal choice of the modulation scheme. Indeed, the demodulator processes only 44.4% of a received 2-D symbol and the space between 2-D symbols, although necessary to avoid ISI, was not optimized. Therefore, there is clearly room for improving the proposed multilevel 2-D bar codes in terms of both BER and storage rate. For example, we obtained a 64.5% increase in storage rate by using the following modulation parameters: $6 \times 6$ pixel 2-D symbols, 1 pixel of inter-symbol space, and $r_{im} = 600$ ppi. This scheme uses the same physical size for the 2-D symbols but reduces the physical inter-symbol space to a third of the previous one (see Figure 2.18). In Table 2.8, we report the storage capacities (at zero BER) of two multilevel 2-D bar codes employing this modulation scheme. For comparison purposes, we also show the public domain 2-D bar codes with the highest storage capacities. Surprisingly, these results were obtained by keeping the same constellations shown in Table 2.2 and only by recomputing the channel model parameters in Tables 2.3a
### Table 2.7: BER results of multilevel 2-D bar codes: (a) $p_1 - s_1$, (b) $p_5 - s_1$ ($N = 2048$).

| $R$ | bytes / in$^2$ | $Z \sim \mathcal{N}(0, \sigma_Z^2)$ | $(Z|X = x) \sim \mathcal{N}(0, \sigma_{Z|X}(x))$ | $(Z|X = x) \sim \mathcal{GGD}(0, \sigma_{Z|X}(x), \gamma_{Z|X}(x))$ |
|-----|----------------|------------------------------------|----------------------------------|--------------------------------------------------|
| 2.4 | 1346           | 0                                  | 0                               | 0                                                |
|     | 1400           | 1.909 × 10$^{-3}$                 | 8.567 × 10$^{-5}$               | 3.671 × 10$^{-5}$                                |
| 2.5 | 1590           | 4.640 × 10$^{-2}$                 | 3.973 × 10$^{-2}$               | 3.930 × 10$^{-2}$                                |
|     | 1684           | 5.012 × 10$^{-2}$                 | 4.146 × 10$^{-2}$               | 4.121 × 10$^{-2}$                                |
| 2.8311 | 1459          | 0                                  | 0                               | 0                                                |
|     | 1515           | 3.392 × 10$^{-5}$                 | 3.392 × 10$^{-5}$               | 1.131 × 10$^{-5}$                                |
| 2.9631 | 1663          | 4.985 × 10$^{-3}$                 | 3.296 × 10$^{-3}$               | 2.874 × 10$^{-3}$                                |
|     | 1684           | 5.229 × 10$^{-3}$                 | 3.855 × 10$^{-3}$               | 3.866 × 10$^{-3}$                                |
and 2.3m. In fact, the improvement in storage rate was due to the used demodulation algorithm which neglects border pixels and thus provides some resistance to ISI.

<table>
<thead>
<tr>
<th>Name</th>
<th>bytes / in²</th>
</tr>
</thead>
<tbody>
<tr>
<td>Data Matrix</td>
<td>1555</td>
</tr>
<tr>
<td>Aztec Code</td>
<td>1888</td>
</tr>
<tr>
<td>QR Code</td>
<td>1941</td>
</tr>
<tr>
<td>Multilevel 2-D Bar Code (p_1 - s_1)</td>
<td>2211</td>
</tr>
<tr>
<td>Multilevel 2-D Bar Code (p_5 - s_1)</td>
<td>2397</td>
</tr>
</tbody>
</table>

Table 2.8: High rate 2-D bar codes.

## 2.6 Conclusions and Related Work

In this Chapter, we highlighted the attractiveness of multilevel 2-D bar codes for applications requiring high rate storage. We have also shown how to apply MLC/MSD, originally developed for the AWGN channel, to the P-S channel in the context of multilevel 2-D bar codes. A key point is the construction of a simplified model of the P-S channel specifically adapted for this application. This model assumes no ISI and perfect synchronization, but it takes into account the dependence between the channel input and the noise. Our approach can be applied to other printing and reading devices as well as to enhance existing B&W 2-D bar codes.

It is worth to mention, that Malvido et al. [29] proposed an extension of the presented P-S channel model by constructing a pixel-level relationship between the input and output of the P-S channel. The main rationale of their work is to exploit as much information as possible from a received noisy 2-D symbol. Although the new model has successfully been tested on a mobile phone and a webcam (assuming small projective distortions), the actual improvement in terms of BER with respect to our model is minor. Indeed, the obtained BERs are of the same order as the ones obtained using our model.

The experimental results in this Chapter show that, in comparison to other public domain 2-D bar codes, the storage capacities of the designed multilevel 2-D bar codes are
the highest. Therefore, we conclude that multilevel 2-D bar codes are very promising candidates to meet the high storage rate requirements of many new multimedia security and data management applications. In particular, for the problem of text document authentication, the fact that multilevel 2-D bar codes can hold an important amount of information within a small surface makes them very attractive because of their low price and the wide availability of readers.

The main results of this Chapter have been published in [101, 102].
3.1 Introduction

The early methods of document authentication based on data hiding embed the document’s identifier as secret data. However, the storage of the document’s identifier is insecure by itself. Indeed, the hidden data is vulnerable to the *copy attack* [30, 31], which estimates the modulated features from the protected document (without knowing the secret key $K_{DH}$) and remodulates another document identically, thus creating an ambiguity [32, 33]. Therefore, a document should be authenticated using, at the same time, content dependent data, such as a hash or a digital signature, and content independent data, such as the document’s identifier.

We are interested in a scenario where the documents are stored and distributed in electronic form as well as in printed form. Such a scenario is typical for the Internet, research and educational institutions, governmental bodies, small and medium businesses, etc.

There are two possible designs of a document authentication system based on digital data hiding. The first design assumes that the payload (hidden message) is content-independent and fixed for all documents or randomly generated from a user key [34]. Content authentication is performed by deciding whether the hidden message is present or not in the document under analysis. However, this approach poses a lot of security concerns for both spread-spectrum and quantization-based data hiding techniques because the secret information (the hidden message, the key, or both) can be estimated and then a fake document can be remodulated with this information similarly to the above mentioned copy attack. This is the so-called *remodulation attack* [30, 31]. In the second design, the payload is assumed to be content dependent [35, 36]. This prevents the use of copy or remodulation attacks. For this reason, we will concentrate on the second design.

Although the approach we follow is also based on robust hashing and data hiding, there are several important differences with [35, 36]. First, we propose a general information-theoretic formulation of the authentication problem, whereas the above approaches mostly advocate practical frameworks. Second, the major difference with the work in [35] consists in the definition of authentication security. In [35], the security of the system was
solely expressed in terms of the security of the hash based on the definition of its entropy. This has two important issues: (a) the security leakage of the data hiding subsystem was not taken into account in the security analysis of the complete system, and (b) since the security of the hash was solely defined based on the entropy of the secret features, the security leakage resulting from the actual tamper-protected document was disregarded. Last, Fei et al. [36] considered only key-dependent hashing and did not analyze the security of the data hiding subsystem, i.e. a key was not defined for the data hiding block in Figure 1 of the cited paper. Moreover, depending on the application the forger can learn more observing multiple documents, pages or paragraphs protected with the same secret key, or, potentially, by observing the same document protected with different keys. However, in order to be compliant with the above mentioned publications, we will also constrain our analysis to the case of a single document.

This Chapter has the following structure. The basic theoretical formulation of the authentication problem together with possible solutions are considered in Section 3.2. In Section 3.3, we formally define an authentication system based on data hiding and robust hashing and demonstrate its suitability for dealing with the authentication problem. Section 3.4 contains the security analysis of a document authentication system based on robust hashing and data hiding. Finally, Section 3.5 concludes the Chapter.

3.2 The General Document Authentication Problem

In this Section, we consider the problem of authenticating a document from an information-theoretic point of view. The generic block diagram of an authentication system meant to solve this problem is shown in Figure 3.1.

![Block diagram of a generic document authentication system.](image)

According to the presented setup, the encoder $\phi$ is under the control of the tamper-proofer who has access to the original document $X^N \in \mathcal{X}^N$, and to the uniquely assigned secret key $K \in \mathcal{K}$. The original document $X^N$ is assumed to be distributed according to $p(x^N)$ and the key $K$ according to a uniform distribution over $K$. The original document $X^N$ and the key $K$ are used at the encoder $\phi$ to generate a tamper-resistant document $Y^N \in \mathcal{Y}^N$, which is semantically equal and visually identical to $X^N$ but is sensitive to malicious modifications. The tamper-resistant document $Y^N$ is communicated through a channel, described by the conditional probability distribution $p(v^N | y^N)$, which introduces some legitimate distortions to $Y^N$. The resulting distorted document is denoted by $V^N \in \mathcal{V}^N$. The decoder $\xi$, which is under the control of the authenticator, makes a decision about the authenticity of $V^N$ based on the knowledge of the secret key $K$. Thus, a
3.2 The General Document Authentication Problem

document authentication system consists of the set \( \{X^N, p(x^N), \mathcal{K}, \mathcal{Y}^N, p(v^N|y^N), V^N\} \), the encoder’s allowable distortion \( D^E \) between \( X^N \) and \( Y^N \), the channel’s allowable distortion \( D^C \) between \( Y^N \) and \( V^N \), and the encoder-decoder pair:

\[
\varphi : X^N \times \mathcal{K} \rightarrow \mathcal{Y}^N, \\
\xi : \mathcal{Y}^N \times \mathcal{K} \rightarrow \mathcal{B} = \{0, 1\}.
\]

The problem of deciding whether the received document \( V^N \) is authentic or not can be considered as a binary hypothesis testing problem [37]. One can assume that \( H_0 \) corresponds to the hypothesis that the received document is authentic, and \( H_1 \) to the alternative hypothesis (i.e. the received document is fake). Therefore, the task of the authentication system is to decide which of the two hypotheses should be accepted given the realizations of \( V^N \) and \( K \), namely \( v^N \) and \( k \). This general idea was first suggested by Maurer [37] and extended to steganographic applications by Cachin [38] and more recently elaborated by Wang and Moulin [39]. In document authentication, this approach was considered in [35, 36].

We will assume that the hypothesis test problem can be stated as:

\[
\begin{align*}
H_0 &: (V^N|K = k) \sim p^0(v^N|k), \\
H_1 &: (V^N|K = k) \sim p^1(v^N|k),
\end{align*}
\]

where \( H_0 \) corresponds to the decision 0 and \( H_1 \) to 1.

Various tests can be performed, for example the Bayesian, minimax or Neyman-Pearson tests; however, we will use the optimal Neyman-Pearson test in our formulation due to the particularities of the authentication problem, which are discussed below.

Disregarding the chosen testing strategy, two types of errors are possible: a type I error or a false alarm occurs, if an authentic document is decided to be a fake, and a type II error or a miss occurs, if a fake document is considered to be authentic.

According to the Neyman-Pearson test, the goal of the authentication system designer is to keep the probability \( P_F \) of false alarm fixed and to minimize the probability \( P_M \) of missing a fake document. Contrarily, the objective of the forger is to maximize \( P_M \) by modifying the tamper resistant document \( Y^N \) in such a way that the malicious modifications are not detected by the decoder. These conflicting objectives can be formulated as a game between the authentication system designer and the forger:

\[
\min_{\varphi, \xi} \max_{p(v^N|y^N)} P_M(\varphi, \xi, p(v^N|y^N)),
\]

which depends on the particular encoder/decoder pair \( \varphi, \xi \) and the channel’s conditional probability distribution \( p(v^N|y^N) \) for a fixed \( P_F \).

Given the secret key \( k \), the Neyman-Pearson test states that for a given maximal tolerable probability \( P_F(k) \) of false alarm, the probability of miss authentication \( P_M(k) \) can be minimized by accepting hypothesis \( H_0 \) if, and only if, the log-likelihood ratio defined as:

\[
\ell(v^N|k) \triangleq \log_2 \frac{p^0(v^N|k)}{p^1(v^N|k)}, \tag{3.1}
\]

satisfies:

\[
\ell(v^N|k) \geq T,
\]
for some threshold \( T \).

We define the probabilities of false alarm and miss for a given key \( k \) as:

\[
P_F(k) = \Pr\{\ell(V^N|k) < T|H_0 \text{ true}\},
\]

and:

\[
P_M(k) = \Pr\{\ell(V^N|k) \geq T|H_1 \text{ true}\},
\]

respectively.

The conditional relative entropy \( D(p^1_{V^N|K}||p^0_{V^N|K}) \), defined as the expected value of the log-likelihood function in (3.1) with respect to \( p^1(V^N, k) \), measures the level of distinguishability between the two involved distributions:

\[
D(p^1_{V^N|K}||p^0_{V^N|K}) = \sum_{k \in K} p_K(k) \sum_{v^N \in V^N} p^1(v^N|k) \log_2 \frac{p^1(v^N|k)}{p^0(v^N|k)},
\]

where \( p(k) \) is the distribution of \( K \) on \( K \), i.e. \( p(k) = 1/|K| \), for all \( k \in K \).

In this case, the average probabilities of false alarm and miss, \( P_F = \sum_{k \in K} p(k) P_F(k) \) and \( P_M = \sum_{k \in K} p(k) P_M(k) \), respectively, satisfy [37]:

\[
P_M \log_2 \frac{P_M}{1 - P_F} + (1 - P_M) \log_2 \frac{1 - P_M}{P_F} \leq D(p^1_{V^N|K}||p^0_{V^N|K}).
\]

Fixing \( P_M = 0 \), one can obtain a lower bound on the average probability of false alarm:

\[
P_F \geq 2^{-D(p^1_{V^N|K}||p^0_{V^N|K})}.
\]

The complete system performance limits, in terms of error exponents, can be obtained according to Stein’s lemma [26, Theorem 12.8.1]:

\[
P_M \approx 2^{-D(p^0_{V^N|K}||p^1_{V^N|K})}, \text{ for a fixed } P_F,
\]

\[
P_F \approx 2^{-D(p^1_{V^N|K}||p^0_{V^N|K})}, \text{ for a fixed } P_M.
\]

The above document authentication system can be considered in the scope of a robust hashing and data hiding (RH-DH) problem. There are several possible system designs based on separation or joint paradigms by analogy with Shannon’s source-channel communication problem [26]. Here, one faces the same problems of performance optimality, complexity as well as the additional issue of security. Since the optimal system structure still remains an open theoretical problem, we analyze the possible advantages of both approaches.

**Separation approach:** by analogy with Shannon’s separation principle, one can assume that the RH-DH problem can be readily separated into robust hashing and data hiding parts. This approach is schematically presented in Figure 3.2. The fact that the system can be separated in such a way without sacrificing optimality is conceptually and practically plausible. However, the analogy would not be complete without mentioning some differences. Shannon’s source-channel separation theorem assumes that the source coding part does not require any knowledge about the channel statistics to produce the input for
the channel coding part. Similarly, for the channel coding part, the structure of the source code is irrelevant too. Contrarily, the robust hashing part should take into account the possible legitimate channel distortions to produce a robust hash. This very important fact makes also a significant difference between traditional cryptographic hashing and hashing for multimedia authentication problems that involve electronic images, video, audio, and text.

**Joint approach:** while the separation approach provides a pragmatic solution to the classical source-channel communication problem, it does not claim to be unique. In fact, this approach is very expensive in terms of delay and complexity. That is why less expensive solutions can be found, which abandon the separation principle and utilize joint source-channel coding. An extreme case of this approach is well-known as uncoded transmission [40]. A similar consideration might be valid for the document authentication problem where a joint optimal approach could be suggested similarly to the setup shown in Figure 3.1, where no particular form of separation is assumed. We consider this approach as a subject of future research concentrating in this Chapter on the separation approach.

**Rate matching:** by analogy with a source coding and channel coding scheme suggested by the separation principle, one can consider the document authentication problem based on separate robust hashing and data hiding as a rate-matching problem. According to this interpretation, one should match, from one side, the rate of the hash code, selected to satisfy all the authentication, security and robustness requirements, and, from the other, the rate of reliable hidden data communications. Contrary to the source coding and channel coding scheme where the performance criterion is the distortion of the source for a given channel, in the RH-DH problem one has to deal with $P_M$ for a fixed $P_F$, a given legitimate channel, and a set of various possible counterfeiting attacks. Moreover, one should also take into account possible security leakages of this scheme that can be efficiently used by the forger to produce a fake document using quite involved attacks with low time-complexity.

### 3.3 Document Authentication Based on Robust Hashing and Data Hiding

In this Section, we consider in detail all the elements of an authentication system based on the separation of robust hashing and data hiding as shown in Figure 3.2 and indicate the conditions of rate matching. According to this setup, the tamper-proofer has access to the original document $X^N \in \mathcal{X}^N$, and the uniquely assigned secret keys $K_H$ and $K_{DH}$, that are assumed to be uniformly distributed over the sets $\mathcal{K}_H = \{1, 2, \ldots, |\mathcal{K}_H|\}$ and $\mathcal{K}_{DH} = \{1, 2, \ldots, |\mathcal{K}_{DH}|\}$, for robust hashing and data hiding, respectively. We also assume that $X^N$ is distributed according to $p(x^N)$ and that:

$$p(x^N) = \prod_{i=1}^{N} p_X(x_i).$$

The original document $X^N$ and the secret key $K_H$ are used to generate the robust hash $L \in \mathcal{L}$. The robust hash $L$ is then mapped to the hash message $M \in \mathcal{M}$. The hash message
3 Information-Theoretic Framework for Document Authentication

Figure 3.2: Document authentication based on the separation of robust hashing and data hiding.
3.3 Authentication via Robust Hashing and Data Hiding

$M$ is next encoded into the watermark $W^N \in \mathcal{W}^N$ based on $X^N$ and the secret key $K_{DH}$. The watermark $W^N$ is embedded into the original document $X^N$, resulting in the tamper-resistant document $Y^N \in \mathcal{Y}^N$. The tamper-resistant document $Y^N$ is communicated through the channel $p(v^N|y^N)$, which introduces some legitimate distortions. For the authentication task, the authenticator has access to the distorted document $V^N \in \mathcal{V}^N$, and the secret keys $K_H$ and $K_{DH}$. The procedure utilized by the authenticator is the following. The decoder outputs $\hat{M}$, an estimate of $M$, based on the distorted document $V^N$ and the secret key $K_{DH}$. Additionally, the hash $\hat{L}$ is computed from $V^N$ and the secret key $K_H$. Finally, a decision about the authenticity of $V^N$ is made based on the comparison of $\hat{M}$ with $\hat{L}$.

We assume that the hash $L$ and the hash message $M$ are uniformly distributed over the sets $\mathcal{L} = \{1, 2, \ldots, |\mathcal{L}|\}$ and $\mathcal{M} = \{1, 2, \ldots, |\mathcal{M}|\}$, respectively. We also assume that $|\mathcal{L}| = 2^{N R_H}$ and $|\mathcal{M}| = 2^{N R_{DH}}$, where $R_H$ and $R_{DH}$ are the hashing and data hiding rates, respectively, and $N$ is the length of all involved vectors, namely $X^N$, $W^N$, $Y^N$ and $V^N$.

**Definition 3.1.** The channel distortion $d_C(y^N, v^N)$ between $y^N \in \mathcal{Y}^N$ and $v^N \in \mathcal{V}^N$ is defined as:

$$d_C(y^N, v^N) \triangleq \frac{1}{N} \sum_{i=1}^{N} d_C(y_i, v_i),$$

where $d_C(y_i, v_i) : \mathcal{Y} \times \mathcal{V} \rightarrow \mathbb{R}^+$ denotes the element-wise distortion between $y_i$ and $v_i$.

**Definition 3.2.** A Gel’fand-Pinsker (G-P) [41] data hiding channel (see Figure 3.3) consists of four alphabets $\mathcal{X}$, $\mathcal{W}$, $\mathcal{Y}$ and $\mathcal{V}$, and a probability transition matrix $p(v^N|w^N, x^N)$, which corresponds to the covert communication of the watermark $W^N$ through the host document $X^N$, represented by the channel $p(y^N|w^N, x^N)$, and the legitimate channel $p(v^N|y^N)$:

$$p(v^N|w^N, x^N) = \sum_{y^N \in \mathcal{Y}^N} p(y^N|w^N, x^N)p(v^N|y^N).$$

This channel is assumed to be memoryless. This means that the probability distributions

![Figure 3.3: Discrete memoryless data hiding channel.](image-url)
\( p(v^N | y^N) \) and \( p(v^N | w^N, x^N) \) are given by:

\[
p(v^N | y^N) = \prod_{i=1}^{N} p_{V|Y}(v_i | y_i),
\]

\[
p(v^N | w^N, x^N) = \prod_{i=1}^{N} p_{V|W,X}(v_i | w_i, x_i).
\]

It is assumed that the legitimate channel \( p(v^N | y^N) \) is subject to the distortion constraint \( D^C \):

\[
\sum_{y^N \in \mathcal{Y}^N} \sum_{v^N \in \mathcal{V}^N} d_C^N(y^N, v^N) p(v^N | y^N) p(y^N) \leq D^C.
\]

We understand under legitimate distortions data manipulations that do not modify the visual content of the data. Possible examples include signal processing operations such as lossy compression, addition of noise, change of contrast, etc.; printing and scanning operations; and geometrical operations such as translation, rotation and scaling.

**Definition 3.3.** The encoding distortion \( d_E(x^N, y^N) \) between \( x^N \in \mathcal{X}^N \) and \( y^N \in \mathcal{Y}^N \) is defined as:

\[
d_E(x^N, y^N) \triangleq \frac{1}{N} \sum_{i=1}^{N} d_E(x_i, y_i),
\]

where \( d_E(x_i, y_i) : \mathcal{X} \times \mathcal{Y} \to \mathbb{R}^+ \) denotes the element-wise distortion between \( x_i \) and \( y_i \).

**Definition 3.4.** A \( (2^{NR_DH}, N) \) G-P data hiding code for the data hiding channel consists of a message set \( \mathcal{M} = \{1, 2, \ldots, 2^{NR_DH}\} \), an encoding function:

\[
f : \mathcal{M} \times \mathcal{X}^N \times K_{DH} \to \mathcal{W}^N,
\]

an embedding function:

\[
\psi : \mathcal{W}^N \times \mathcal{X}^N \to \mathcal{Y}^N,
\]

subject to the distortion constraint \( D^E \):

\[
\frac{1}{|K_{DH}| |\mathcal{M}|} \sum_{k_{DH} \in K_{DH}} \sum_{m \in \mathcal{M}} \sum_{x^N \in \mathcal{X}^N} d_E^N(x^N, \psi(f(m, x^N, k_{DH}), x^N)) p_{X^N}(x^N) \leq D^E,
\]

and a decoding function:

\[
g : \mathcal{Y}^N \times K_{DH} \to \mathcal{M}.
\]

**Definition 3.5.** We define the average probability of error for a \( (2^{NR_DH}, N) \) G-P data hiding code as:

\[
P_E^{(N)} = \frac{1}{|K_{DH}| |\mathcal{M}|} \sum_{k_{DH} \in K_{DH}} \sum_{m \in \mathcal{M}} \Pr\{g(V^N, k_{DH}) \neq m | K_{DH} = k_{DH}, M = m\}.
\]

**Definition 3.6.** A rate \( R_{DH} = \frac{1}{N} \log_2 |\mathcal{M}| \) is achievable for distortions \( (D^E, D^C) \), if there exists a sequence of \( (2^{NR_DH}, N) \) data hiding codes such that \( P_E^{(N)} \to 0 \) as \( N \to \infty \).
Definition 3.7. The capacity of the G-P data hiding channel is the supremum of all achievable rates for distortions \((D^E, D^C)\).

Theorem 3.1 (G-P data hiding capacity). A rate \(R_{DH}\) is achievable for the secret key \(K_{DH} = k_{DH}\), the distortion pair \((D^E, D^C)\), and the fixed legitimate channel \(p(v|y)\) if, and only if, \(R_{DH} < C\), where:

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

and \(U\) is an auxiliary random variable distributed over the set \(\mathcal{U}\), with \(|\mathcal{U}| \leq |\mathcal{X}| |\mathcal{W}| + 1\).

The proof of this theorem in the more general form of an active attacker is provided by Moulin and O’Sullivan [42] and the details can be found in the referred paper. However, it is important to emphasize that the main difference with our setup is the codebook construction and the corresponding interpretation of the secret key \(K_{DH}\). In the scope of this thesis, the key \(K_{DH}\) is considered uniquely as an index that defines the codebook of a particular user. Contrarily, Moulin and O’Sullivan have a broader understanding of the key \(K_{DH}\) as a sort of side information shared between the encoder and the decoder, where \(K_{DH}\) can be in some relationship with \(X^N\). Therefore, we assume that \(K_{DH}\) is solely a cryptographic key that is independent of \(X^N\).

The codebook construction in the proof of the achievable part of this theorem uses a random binning technique and the concept of strong joint typicality [26]. The main idea is to trade-off the number of codewords needed at the encoder in each message bin in order to cancel the interference \(X^N\) and the number of uniquely distinguishable codewords at the decoder. In the following three paragraphs, we assume that \(K_{DH} = k_{DH}\).

Codebook construction: introduce an auxiliary random variable \(U\) with alphabet \(\mathcal{U}\) via \(p(u|x)\). Generate \(|\mathcal{J}| |\mathcal{M}|\) codewords \(u^N(m, j, k_{DH})\), where \(m \in \mathcal{M}\), \(j \in \mathcal{J} = \{1, 2, \ldots, 2^{NP_{DH}}\}\) and \(k_{DH} \in \mathcal{K}_{DH}\), independently at random according to the marginal distribution \(p(u)\). The number \(NP_{DH}\) can be interpreted as the number of bits used to represent the interference \(X^N\). The codebook is organized as shown in Figure 3.4.

Encoder: (see Figure 3.5) given the message \(m\), the interference \(x^N\), and the secret key \(k_{DH}\), the encoder seeks a codeword \(u^N(m, j, k_{DH})\) such that \((u^N(m, j, k_{DH}), x^N) \in \mathcal{A}^*(N)(U, X)\), i.e. the encoder seeks a jointly strongly typical pair \((u^N(m, j, k_{DH}), x^N)\) in the set of jointly strongly typical sequences \(\mathcal{A}^*(N)(U, X)\). Therefore, the message \(m\) defines the bin and the interference \(x^N\) is exploited to select a particular codeword \(u^N(m, j, k_{DH})\) from this bin (see Figure 3.4). If such a codeword is found, the encoder produces \(w^N\) according to a deterministic mapping \(w^N = \zeta(u^N(m, j, k_{DH}), x^N)\).

Decoder: (see Figure 3.6) given the channel output \(v^N\) and the secret key \(k_{DH}\), the decoder seeks a codeword \(w^N(m, j, k_{DH})\) such that \((w^N(m, j, k_{DH}), v^N) \in \mathcal{A}^*(N)(U, V)\) in the set of all \(|\mathcal{J}| |\mathcal{M}|\) codewords. If the decoder finds a unique jointly strongly typical pair, then it declares that the sent message was \(\hat{m} = m\). Otherwise, an error is declared.

We now concentrate on robust hashing. Without loss of generality, we assume that \(\mathcal{X} = \mathcal{Y} = \mathcal{V}\).

Definition 3.8. The distortion \(d(x_1^N, x_2^N)\) between \(x_1^N \in \mathcal{X}^N\) and \(x_2^N \in \mathcal{X}^N\) is defined as:

\[
d(x_1^N, x_2^N) \triangleq \frac{1}{N} \sum_{i=1}^{N} d(x_{1i}, x_{2i}),
\]
where \( d(x_{1i}, x_{2i}) : \mathcal{X} \times \mathcal{X} \rightarrow \mathbb{R}^+ \) denotes the element-wise distortion between \( x_{1i} \) and \( x_{2i} \).

**Definition 3.9.** A \((2^{N_{RH}}, N)\) robust hash code consists of a robust hash set \( \mathcal{L} = \{1, 2, \ldots, 2^{N_{RH}}\} \) and a robust hash function:

\[
\theta : \mathcal{X}^N \times \mathcal{K}_H \rightarrow \mathcal{L}.
\]

The construction of a robust hash code should satisfy several conflicting requirements. The robust hash function produces a secure robust hash \( L \in \mathcal{L} \) given \( X^N \) and \( K_H \). Contrary to classical cryptographic hashing, where two documents that differ in only a single bit have different and ideally independent hashes, we require that two documents \( X_1^N \) and \( X_2^N \) that are perceived (respectively, understood) by the observer (respectively, by the
3.3 Authentication via Robust Hashing and Data Hiding

\[
\hat{m} = 1 \left( \begin{array}{c} \begin{pmatrix} u^N(1,1,k_{DH}) \\ u^N(1,2,k_{DH}) \\ \vdots \\ u^N(1,J,k_{DH}) \end{pmatrix} \\ \vdots \\ \begin{pmatrix} u^N(|\mathcal{M}|,1,k_{DH}) \\ u^N(|\mathcal{M}|,2,k_{DH}) \\ \vdots \\ u^N(|\mathcal{M}|,|\mathcal{J}|,k_{DH}) \end{pmatrix} \end{array} \right) v^N
\]

Figure 3.6: Gel’fand-Pinsker decoder.

reader) to be similar, in a well defined perceptual sense, have the same or almost the same hash, even if \(X_1^N\) and \(X_2^N\) have small bit-level discrepancies. In practice, it also means that if \(X_2^N\) is obtained via a mapping \(p(x_2^N|x_1^N)\) of \(X_1^N\) such that \(E[u^N(X_1^N,X_2^N)] \leq D\), i.e. the dissimilarity between \(X_1\) and \(X_2\) is below a predefined value of maximum allowable variation \(D\), one should expect \(\theta(X_1^N,K_H) \approx \theta(X_2^N,K_H)\). Additionally, the robust hash function should be secure in the sense that having a document \(X^N\), the forger cannot generate the corresponding robust hash without the knowledge of the key \(K_H\). At this moment, we address neither the issue of computational complexity of the robust hash function nor its collision resistance.

Rather than follow some particular design of a robust hash function which is generally the case for most of the publications on this subject [35, 36, 43], we will consider a generic approach. We will assume that robust hashing is accomplished in some transform domain achieved by applying a transform \(\mathcal{T} : \mathcal{X}^N \times K_H \rightarrow \mathcal{X}'^N\) that is key-dependent in the general case. We will denote the transformed vector of length \(N\) by \(\hat{X}^N = \mathcal{T}(X^N, K_H)\), where \(\hat{X}^N \in \mathcal{X}'^N\) as shown in Figure 3.7.

Figure 3.7: Generic approach to robust hashing.

The application of the transform \(\mathcal{T}\), besides increasing the security, is additionally supposed to provide the hash function robustness against various legitimate degradations within \(D\). The robustness or invariance to legitimate distortions of the vector coefficients \(\hat{X}^N\) might be achieved in two different ways: (a) by using an invariant domain [43, 44,
or (b) by using robust feature quantization [46, 47] as represented by the $Q$ block in Figure 3.7.

We present a random code construction for a robust hash code and analyze its performance assuming that the legitimate distortion $D$ is the same in the transform domain. Although it is not true in the general case, it can be assumed true for all orthogonal transforms. Suppose we choose a mapping $p(\hat{x} | x)$ and compute $p(\hat{x})$ as the marginal distribution of $p(\hat{x}, \tilde{x}) = p(\hat{x} | \tilde{x})p(\tilde{x})$.

**Robust hash code construction:** first, generate $2^{N_R H}$ codewords $\tilde{x}^{N'}(l, k_H)$, for each $k_H \in \mathcal{K}_H$, by choosing each of the $N'2^{N_R H} | \mathcal{K}_H|$ codeword symbols $\tilde{x}_i(l, k_H)$ independently at random according to $p(\hat{x})$. It can be shown that if the number $2^{N_R H} | \mathcal{K}_H|$ is smaller than $2^{N_R H}(\hat{x})$, one can hope to have a unique set of sequences in each user codebook. Then, assign indices $l$ to the codewords $\tilde{x}^{N'}(l, k_H)$, $l = 1, 2, \ldots, 2^{N_R H}$, for each $k_H$ in such a way that the Hamming distance between the binary representation of indices $l \equiv b^3(l) = (b_1(l), \ldots, b_S(l))$, $b_s(l) \in \mathcal{B} = \{0, 1\}$, $s = 1, \ldots, S$, for original and legitimately distorted documents is minimal. In order to accomplish this task, one can consider the equivalent problem of robust labeling in the design of channel codes based on bit interleaved coding modulation [48] where Gray labeling is preferred to a random label assignment. The above technique used to construct the codebooks for $| \mathcal{K}_H |$ users can also be considered as a binning technique.

**Robust hash function:** given $x^N$ and $k_H$ or, equivalently, $\tilde{x}^{N'}$ and $k_H$ after transformation via $\Xi$, try to find a codeword $\tilde{x}^{N'}(l, k_H)$ such that $(\tilde{x}^{N'}, \tilde{x}^{N'}(l, k_H)) \in A_{\Xi}(N')(\hat{X}, \hat{X})$, i.e. the two vectors are jointly strongly typical [26]. If one finds such a jointly strongly typical pair, declare the corresponding hash code to be $l$. Otherwise, let $l = 1$.

One can show that the achievable rate $R_H$ of the above robust hash code satisfies:

$$R_H(D) = \min_{\eta} \min_{p(\hat{x} | \tilde{x})} \frac{1}{E[|d(\hat{x}, \tilde{x})|] \leq D} I(\hat{X} ; \tilde{x}^N),$$

if we allow the collision of all vectors in the range of $D$ using a proof similar to the classical Shannon’s source coding theorem [26]. The sketch of the achievability part of the proof is based on the analysis of the bound on average distortion for three different cases: (a) $\tilde{x}^{N'} \notin A_{\Xi}(N')(\hat{X})$, (b) $\tilde{x}^{N'} \in A_{\Xi}(N')(\hat{X})$ but none of the $\tilde{x}^{N'}(l, k_H)$ satisfies $(\tilde{x}^{N'}, \tilde{x}^{N'}(l, k_H)) \in A_{\Xi}(N')(\hat{X}, \hat{X})$, and (c) $\tilde{x}^{N'} \in A_{\Xi}(N')(\hat{X})$ and we find a $\tilde{x}^{N'}(l, k_H)$ with $(\tilde{x}^{N'}, \tilde{x}^{N'}(l, k_H)) \in A_{\Xi}(N')(\hat{X}, \hat{X})$. One can show that in all these cases the distortion will not exceed $D$ with high probability as long as $N' \rightarrow \infty$.

In practice, such a robust hash function can be implemented by using a keyed vector quantizer $Q : \hat{X}^{N'} \times \mathcal{K}_H \rightarrow \mathcal{L}$ as shown in Figure 3.7, where the reconstruction points are selected properly for each key $k_H \in \mathcal{K}_H$ to satisfy the distortion constraint. To simplify the implementation and analysis, one can design a dithered quantizer $Q_0(\tilde{x}^{N'} - d^{N'}_{k_H}) + d^{N'}_{k_H}$, which uses a fixed vector quantizer $Q_0 : \hat{X}^{N'} \rightarrow \hat{X}^{N'}$ in all cases and a length-$N'$ dither vector $d^{N'}_{k_H}$ generated from the secret key $k_H$. Further simplification can be achieved using uniform scalar quantization instead of vector quantization.

**Definition 3.10.** A decision function is defined by a mapping:

$$\eta : \mathcal{L} \times \mathcal{M} \rightarrow \mathcal{B} = \{0, 1\},$$
that takes two hashes to produce a binary decision.

The performance of the overall document authentication system is measured in terms of the average probability of miss \( P_M \) for a given maximal tolerable average probability of false alarm \( P_F \), similarly to the results presented in Section 3.2. The authentication task amounts to select one of the two hypothesis \( \{ H_0, H_1 \} \) based on the binary representations of the hash computed from the observed data \( V^N \) and that from the decoded message, i.e. \( \hat{L} \equiv B^S(\hat{L}) \) and \( \hat{M} \equiv B^S(\hat{M}) \), respectively. The binary decision is taken by comparison of the number of different bits between \( B^S(\hat{L}) \) and \( B^S(\hat{M}) \) with respect to a predefined threshold \( T(P_F) \).

The main issue in the information-theoretic analysis of the studied setup consists in the derivation of direct and converse theorems for reliable document authentication under the assumed types of legitimate distortions. Hereafter, we concentrate on the direct theorem, the converse still remaining an open theoretical problem.

**Theorem 3.2** (authentication based on robust hashing and data hiding). If \( X^N \) is a finite alphabet stochastic process that satisfies the asymptotic equipartition property (AEP) [26], then there are robust hashing and data hiding codes with specified average probability of false alarm \( P_F \) and vanishing average probability of miss \( P_M \) as \( N \to \infty \), if the rates of the hashing and data hiding codes \( R_H \) and \( R_{DH} \) satisfy \( R_H \leq R_{DH} < C \).

The proof tries to establish the information-theoretic bounds on the rate \( R_H \) of the robust hash code providing reliable document authentication for the legitimate channel and the assumed class of malicious attacks, as well as on the rate of reliable hidden communications \( R_{DH} \leq C \), where \( C \) is the G-P data hiding capacity for the data hiding channel. Here, we sketch the proof of Theorem 3.2. Assuming that the encoder and decoder share the same pair of secret keys \( k_H \) and \( k_{DH} \), we will use three quantities defining the performance of the authentication system, namely \( P_E^{(N)} \), \( P_F \) and \( P_M \). Since we assume that \( X^N \) satisfies the AEP, the robust hash code with parameters specified in Definition 3.9 can be constructed with \( 2^{NR_H} \) codewords for each key \( k_H \), i.e. with rate \( R_H \). According to the considered robust labeling, the index of the hash for a given key \( k_H \) is fed to the input of the channel encoder by assigning \( m = l \). The encoder maps this index with the key \( k_{DH} \) into the sequence \( Y^N \) sent to the decoder according to Definition 3.4 with rate \( R_{DH} \) that satisfies the conditions of Theorem 3.1.

One can transmit the hash index \( m \) to the decoder with probability of error \( P_E^{(N)} \) less than \( \epsilon \), if:

\[
R_H + \epsilon = R_{DH} < C.
\]

The decoder correctly estimates the sent message \( \hat{m} \) with high probability, if \( R_{DH} < C \) according to Theorem 3.1. Finally, assuming the robust hash code is able to reliably extract the index \( \hat{l} \) from the channel output \( V^N \) under a specified allowable legitimate distortion \( D \) and assuming the knowledge of the secret key \( k_{DH} \), one can guarantee the authenticity of \( V^N \) with specified \( P_F \) for a given rate \( R_H \) with vanishing average probability of miss \( P_M \).
3.4 Security Analysis of Document Authentication

In this Section, we consider the proposed authentication system from a security perspective. As was pointed out in Section 3.2, the overall objective of a forger is to modify the document in such way that the modification is undetectable, i.e. to increase the probability $P_M$ by operating in the range of legitimate distortions $D_C$. Contrary to various attacking scenarios that, in the general case, assume the availability of several copies of protected documents $Y_1^N, Y_2^N, \cdots, Y_j^N$ protected with the same key, which have been intensively studied in the publications on authentication summarized by Maurer [37], we will focus the case of a single available document $Y^N$ leaving the extension of our framework to multiple documents for future study.

To fully benefit from the available copy of $Y^N$, we recall Shannon’s equivocation concept considered in [49, 50, 51] and originally formulated with respect to the parameters of the scheme such as the secret key, the message, the host, the auxiliary random variable $U^N$ and the hash $L$.

We consider the security analysis of the authentication system as a further development of the reversibility principles introduced in [52]. Here, we adapt the security framework in [52] to the authentication application by defining security as the amount of trial efforts required to reveal the secret information in order to design the worst forgery strategy in the sense that the desired document alteration is unnoticeable by the authentication system, i.e. maximize $P_M$ by combining all public information with the revealed secret one while keeping the introduced distortions in the allowable ranges. The trial efforts are considered in a broad sense for the authentication applications. In the scope of this Section, we will only consider a particular aspect of complexity, which is expressed by the number of checks to be performed to reveal some secret or to attack the scheme. This definition of security is also coherent with the definitions given in previous publications [49, 50, 51].

In fact, we show that the knowledge of $U^N = u^N(m, j, k_{DH})$ and $\hat{X}^N = \hat{x}^N(l, k_H)$ is sufficient to achieve the forger’s goal under certain circumstances. Thus, the amount of efforts in terms of number of trials is understood as the real security of the robust data hiding system. The larger this amount, the higher the security is. It should be also pointed out that the key difference of this approach with previously considered ones consists in the fact that we analyze the security leakage coming from the auxiliary random variable $U$ and $\hat{X}^N$ rather than considering particular messages, keys or hosts. From the information about $U^N$ and $\hat{X}^N$, one can deduce the knowledge of their components, i.e. the key, message and even the host, and apply this knowledge selectively depending on the particular application scenario.

**Lemma 3.1** (equivocation about the data hiding code). The observation of $Y^N$ reduces the ambiguity about $U^N$ from $h(U^N)$ to $h(U^N|Y^N)$:

$$h(U^N|Y^N) = h(U^N) - I(U^N; Y^N).$$

Therefore, the corresponding forger’s complexity in revealing the used $u^N$ in the data hiding code is:

$$O(2^{h(U^N|Y^N)}).$$
This shows that the forger can reduce the volume of his search space from $2^{h(U^N)}$ to $2^{h(U^N|Y^N)}$ trials by performing jointly-typical decoding only with those codewords that are located within some “distance” from the minimum mean-squared error (MMSE) estimate $E[U^N|Y^N]$ of $U^N$. It can be argued that this distance is determined by the estimation variance $\sigma_{U^N|Y^N}^2$, which therefore defines the ambiguity volume $2^{h(U^N|Y^N)}$.

Thus, the forger will perform the search not among all $u^N \in A^s(U^N)$ but only among those $u^N$'s that are within the ambiguity volume which considerably reduces his complexity. Finally, it should be pointed out that the knowledge of $U^N$ provides also information about $\hat{M}$, $J$ and $K_{\text{DH}}$ according to the assumed codebook construction.

Similar conclusions can be deduced concerning the security of the hash code.

**Lemma 3.2** (equivocation about the hash code). The observation of $Y^N$ reduces the ambiguity about $\hat{X}^{N'}$ from $h(\hat{X}^{N'})$ to $h(\hat{X}^{N'}|Y^N)$:

$$h(\hat{X}^{N'}|Y^N) = h(\hat{X}^{N'}) - I(\hat{X}^{N'}; Y^N).$$

Since $\hat{X}^{N'}(L, K_{\text{H}})$ is a function of $L$ and $K_{\text{H}}$, the fact of revealing information about $\hat{X}^{N'}$ implies the availability of information about $L$ and $K_{\text{H}}$. Therefore, the corresponding complexity of the forger in revealing the information about $\hat{X}^{N'}$ used in the hash code is:

$$O(2^{h(\hat{X}^{N'}|Y^N)}).$$

The above considerations concerning the ambiguity volume are also valid here. Moreover, it should be pointed out that the major difference between the presented approach to security evaluation and the approach suggested in [35], which is only based on the entropy of $\hat{X}^{N'}$, is that the later overestimates the ambiguity volume. Such a definition of security disregards a possible security leakage from $Y^N$ that is taken into account in our formulation. Finally, the security of the data hiding and hash codes is formulated using the same theoretical apparatus which makes our approach more general.

Having defined the equivocations about the data hiding and hash codes, we will consider some possible attacking scenarios against a RH-DH document authentication system for two different key management protocols. This analysis is performed according to Kerckhoffs’ principle [53], which assumes that the forger has access to all the particularities of the authentication system besides the secret keys and the communicated messages.

### 3.4.1 The Case of Identical Keys

Within this key management protocol, we assume that $k_{\text{H}} = k_{\text{DH}} = k$ and investigate the possible security of this scheme. This was a quite typical assumption for the majority of early authentication systems.

The corresponding attacking strategy can be summarized as follows:

1. Given $Y^N$, estimate the center of the ambiguity sphere about the sent codeword $E[U^N|Y^N]$ applying an MMSE strategy.
2. Find $\hat{U}^N$ based on the available $Y^N$ performing $2^{h(U^N|Y^N)}$ possible checks.
3. Find $K$ since the knowledge of $\hat{U}^N$ reveals the complete information about $M$ and $K$.

4. Apply reversibility, i.e. estimate $\hat{X}^N = E[X^N | Y^N, \hat{U}^N]$. Under special conditions considered in [52], one can even achieve perfect reversibility $\hat{X}^N = X^N$.

5. Produce a fake copy $X'^N$ of $X^N$ according to the desire of the forger.

6. Compute a new hash $M'$ from the available $X'^N$ and $K$.

7. Perform embedding of $M'$ into $X'^N$ using $K$.

The corresponding complexity of the forger to perform the above operations is:

$$O(2^{h(|U^N|Y^N)})$$

which coincides with (3.4) indicating the important fact that the security of the hash does not play any role in this key management protocol.

The opposite is also valid, in case the forger tries to reveal the information about $K$ from the hash part of the code:

$$O(2^{h(\hat{X}^N | Y^N)})$$

Since, the forger can always choose the strategy with lower complexity, the final estimate for the security of the considered authentication system is:

$$\min \left\{ O(2^{h(U^N | Y^N)}), O(2^{h(\hat{X}^N | Y^N)}) \right\}.$$

### 3.4.2 The Case of Different Keys

A similar attacking strategy can be applied in the case of two different keys, i.e. $K_H \neq K_{DH}$. The forger can first learn the embedded message based on $Y^N$ with complexity:

$$O \left( 2^{h(U^N | Y^N)} \right),$$

and then use the prior knowledge about the message to learn the key $K_H$ for the hash regeneration from the forged data. This attack is similar in spirit to the so-called known message attack [49, 51] and its complexity can be estimated as:

$$O \left( 2^{h(\hat{X}^N | Y^N, L)} \right).$$

Finally, the total efforts of the forger are:

$$O \left( 2^{h(U^N | Y^N)} + 2^{h(\hat{X}^N | Y^N, L)} \right), \hspace{1cm} \text{(3.2)}$$

Contrarily, the forger can first estimate the message that was embedded based on the hash analysis and then apply the above mentioned known message attack to reveal the information about the secret key $K_{DH}$ used for the data hiding part. Knowing two keys, the forger can perform the reversibility, regenerate the hash, and sequentially embed it into the fake document. The total efforts of the forger in this strategy can be estimated as:

$$O \left( 2^{h(\hat{X}^N | Y^N)} + 2^{h(U^N | Y^N, M)} \right). \hspace{1cm} \text{(3.3)}$$

The resulting efforts are equal to the minimum of (3.2) and (3.3).
3.5 Conclusions

In this Chapter, we considered the problem of authentication of electronic and printed documents. We presented an information-theoretic framework for joint robust hashing and data hiding as well as conducted a performance analysis of a such generic authentication system. We also justified the appropriateness of a document authentication system that is based on the separation of robust visual hashing and data hiding. The key idea for this result is the interpretation of robust hashing as a special kind of source coding consisting of two steps: domain transformation and robust feature quantization. Our main result indicates that if the rate of the hash code is not greater than the achievable rate of the data hiding code, then there exist robust hashing and data hiding codes with a specified average probability of false alarm and vanishing average probability of miss as the length of the source document tends to infinity. Furthermore, by using an information equivocation argument, we also conducted a security analysis of this system and proposed brute force attacking strategies with corresponding complexity estimates that, contrarily to existing studies, crucially rely on the information leakage of the document authentication system.

The main results of this Chapter have been published in [103].
4 Practical Methods for Text Document Authentication

4.1 Introduction

We now focus our attention on the main practical concerns of text document authentication via robust hashing and data hiding, namely the limited data embedding rate offered by current text data hiding methods, which is a direct consequence of the imposed strict constraints on the visible degradation and semantics of text documents, and the lack of reliable and secure robust text hashing functions. As discussed in Section 3.3, our aim is to match the information rates of the robust hash and data hiding codes in order to guarantee reliable authentication of documents in the sense of Theorem 3.2.

Thus, the goal of this Chapter is twofold. First, to address the problem of limited data embedding rate of current text data hiding technologies. Second, to study the properties of possibly good candidate methods for robust text hashing.

For the first goal, we introduce the Gel’fand-Pinsker (G-P) text data hiding framework. Based on this framework, we introduce a new semi-fragile text data hiding method that is fully automatable, has a high information embedding rate, is resistant to printing and scanning, and can be applied simultaneously to both electronic and printed text documents. Moreover, we consider the combination of independent text data hiding methods in order to increase the data embedding rate. Such a strategy is again a natural extension of the G-P framework for text data hiding. As for the second goal, we study two text hashing methods in order to establish their suitability for a text document authentication system.

This Chapter is organized as follows. A brief review of the state of the art of text data hiding is given in Section 4.2. The interpretation of the text data hiding problem as a G-P problem is given in Section 4.3. Concrete examples of G-P text data hiding and the combination of independent text data hiding methods are explained in Section 4.4. In Section 4.5, two methods for text hashing are presented. Experimental results concerning a practical implementation of the color index modulation method and the presented text hashing methods are given in Section 4.6. Finally, Section 4.7 concludes this Chapter.
4.2 State of the Art of Text Data Hiding

Four major groups of methods for data hiding in text documents have appeared in the past fifteen years: syntactic methods [54, 55], where the diction or structure of sentences is transformed without significantly altering their meaning; semantic methods [54, 55], where words are replaced by their synonyms, and sentences are transformed via suppression or inclusion of noun phrase coreferences; open space methods [54, 56], where either the inter-line space, the inter-word space or the inter-character space is modulated; and character feature methods [56, 57, 58], where features such as shape, size or location are modified.

However, syntactic and semantic methods are not suitable for all types of documents (e.g. contracts, identification documents, literary texts) and usually need human supervision. Some open space methods such as inter-line space modulation and inter-word space modulation can be automated, are robust against printing and scanning, but have low information embedding rates. On the other hand, inter-character space modulation and existing character feature methods have higher information embedding rates, but are less or not robust at all against printing and scanning. Since automation is very important for practical applications, we will not consider syntactic or semantic methods in this Chapter.

Is the problem of text data hiding too difficult to solve? The answer to this question depends on the application requirements. For example, for copyright protection applications where robust data hiding is required, the answer might be YES, since the attacker can always use optical character recognition (OCR) to rebuild the text document without any trace of hidden data. On the other hand, we advocate that the answer is NO for applications where either semi-fragile or fragile data hiding is required (e.g. authentication and tamper proofing applications).

4.3 Text Data-Hiding as a Gel’fand-Pinsker Problem

We explain now how the text data hiding problem can be considered as a particular instance of the G-P problem (see Figure 4.1 and Theorem 3.1). The text, where some message $M$ is to be hidden, is represented by $X^N$ and called cover text. Each component

![Figure 4.1: Text data hiding as an instance of the Gel’fand-Pinsker problem.](image-url)
\(X_n, n = 1, 2, \ldots, N\), of \(X^N\) represents one character from this text. Here, we define a character as an element from a given language alphabet (for instance, the Latin alphabet \(\{A, B, \ldots, Z\}\)). This alphabet can also contain punctuation characters as well as other special characters\(^1\). To be more precise, we conceive each character \(X_n\) as a data structure consisting of multiple quantifiable component fields (features): shape (geometric definition), location, orientation, size, color, etc. Given the message \(M\) and the cover text \(X^N\), the encoder looks for a jointly strongly typical pair \((U^N, X^N) \in A^*_N(U, X)\). The watermark \(W^N\) is found via a deterministic mapping \(W^N = \zeta(U^N, X^N)\). The influence of the channel \(p(v|w, x)\) is divided in two stages. In the first stage, \(W^N\) and \(X^N\) are combined via a deterministic mapping, defined as the embedder, to produce the stego text \(Y^N = \psi(W^N, X^N)\). In general, \(Y^N\) needs to satisfy a number of requirements, which can be different according to the selected application. In the second stage, \(Y^N\) may suffer from some intentional or unintentional distortions. We denote by \(V^N\) the resulting distorted version of the stego text. Finally, \(V^N\) is fed to the decoder, which tries to obtain an estimate \(\hat{M}\) of message \(M\).

### 4.3.1 Costa’s Problem

Costa considered the G-P problem in the particular case of zero-mean additive white Gaussian interference and zero-mean additive white Gaussian noise (AWGN) [59]:
\[
V = W + X + Z, \quad X \sim \mathcal{N}(0, \sigma_X^2), \quad Z \sim \mathcal{N}(0, \sigma_Z^2).
\]

Given the watermark power constraint \(E[\|W^N\|^2] \leq N\sigma_W^2\), Costa demonstrated that if \(U\) is defined via \(U = W + \alpha X\), where \(W \sim \mathcal{N}(0, \sigma_W^2)\) is independent of \(X\) and \(\alpha = \sigma_W^2 / (\sigma_W^2 + \sigma_Z^2)\), then the capacity of this channel coincides with the capacity of the AWGN channel \(V = W + Z\), i.e.:
\[
C = C_{AWGN} = \frac{1}{2} \log_2 \left(1 + \frac{\sigma_W^2}{\sigma_Z^2}\right).
\]

### 4.3.2 Practical Implementation of Costa’s Problem

It should be noticed that 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 of such coding schemes, the use of structured codebooks instead of random ones has been proposed [60, 61]. These codebooks are designed based on high-rate quantizers providing the independence between \(W\) and \(X\). For example, in the so-called scalar Costa scheme (SCS) [61] the auxiliary random variable \(U\) is approximated by:
\[
U = W + \alpha'X = \alpha'Q_m(X),
\]
where \(Q_m(\cdot)\) is a scalar quantizer for the message \(m\) and \(\alpha'\) is a compensation parameter. This produces a uniformly distributed watermark \(W = U - \alpha'X = \alpha'Q_m(X) - \alpha'X\).

\(^1\)In fact, we show in Sections 4.4 and 4.6.1 that the particular language alphabet is not relevant for the practical implementation of a text data hiding scheme based on the G-P framework.
The resulting stego text is obtained as:

\[ Y = W + X = \alpha' Q_m(X) + (1 - \alpha') X. \]  

(4.1)

To illustrate the G-P text data hiding interpretation, we consider a practical implementation based on the SCS. For this, we select one character feature, e.g., color, and use it as the cover character \( X \) in (4.1). We show in Figure 4.2 the resulting SCS codebook and an illustration of its usage for text data hiding.

![Figure 4.2: SCS codebook for text data hiding \( (N = 1) \).](image)

Distortion-Compensated Quantization Index Modulation [60] can be seen as a generalization of SCS to the vector case, where \( N \)-dimensional vector quantizers are used instead of scalar ones. For this scheme, which is also well suited for text data hiding, we have:

\[ U^N = W^N + \alpha' X^N = \alpha' Q_m^N(X^N), \]

and the stego text is given by:

\[ Y^N = W^N + X^N = \alpha' Q_m^N(X^N) + (1 - \alpha') X^N. \]  

(4.2)

In our interpretation, this accounts for taking groups of characters in order to build a suitable codebook. For example, for \( N = 8 \), such a codebook would contain the entry \( Q_m^8(\text{TRABAJAR}) \), corresponding to message \( m \) and the group of characters \( x^8 = \text{TRABAJAR} \).

In this work, we fix the decoder to be the minimum Euclidean distance decoder, defined by:

\[ \hat{m}(v^N) = \arg \min_{m \in M} \| v^N - Q_m^N(v^N) \|, \]

where \( v^N \) is the noisy stego text.

In Section 4.4, we give concrete examples on how to apply these methods to the problem of text data hiding. In particular, we show that all previously proposed character feature methods [56, 57, 58], including inter-character space modulation, can be considered as particular cases of the scheme described by (4.1), and that all previously proposed
open space methods [54, 56] modifying either inter-word or inter-line space are all particular cases of the more general scheme described by (4.2), where the exploited character feature is location.

### 4.4 Practical Implementation of Gel’fand-Pinsker Text Data-Hiding

In text authentication and text tamper proofing applications, an alteration of the hidden data indicates that the text document has been modified. Thus, either fragile or semi-fragile text data hiding is needed. A fragile method, intolerant to any modification, would only address electronic documents, whereas a semi-fragile one, robust to some legitimate distortions such as those resulting from print-scan (P-S) processes, would address both electronic and printed documents [62].

In our view, the main requirements for a semi-fragile text data hiding method should be the following:

**R1.** It should work for text documents both in electronic and printed forms.

**R2.** It should be independent of the document’s electronic format provided that this format supports a reasonable level of text description. Modern electronic formats satisfying this condition are: Microsoft’s Word document format (DOC) and rich text format (RTF), Adobe’s PostScript (PS) and portable document format (PDF), \( \LaTeX \) (TEX), and many others.

**R3.** Marked electronic text documents should be converted from one format to another retaining the hidden information in a transparent manner for the end user.

**R4.** Marked text documents should be perceptually undistinguishable from the original versions.

**R5.** It should have a high information embedding rate. Even a single page of text should be sufficient to hold some basic information. For example, the author’s name, time and date of creation, comments, etc.

**R6.** It should be easy to automate. Automation and unsupervised processing are very important features to make the solution attractive for practical applications.

We now illustrate the G-P text data hiding framework developed in Section 4.3 by explaining three practical text data hiding methods. We will also deal with possible ways to combine these methods as well as to increase their robustness.

#### 4.4.1 Color Index Modulation (CIM)

In this original method [63], the stego text is obtained via (4.1), for \( \alpha' = 1 \) and defining the character feature \( X \) to be color (see also Figure 4.2). The main idea of this method is to quantize the color of each character in such a manner that the human visual system (HVS) is not able to distinguish between the original and quantized characters,
but it is still possible for a specialized reader, e.g. a high dynamic range scanner in the case of printed documents. It should also be mentioned that when halftoning is used as printing technology, halftone patterns are used to represent colors. Since the choice of these halftone patterns is not unique (see Section 4.4.3), i.e. there are various halftone patterns reproducing the same color, it is possible to exploit this characteristic to increase the data embedding rate.

An example illustrating this method is shown in Fig 4.3. Therein, the original piece of text uses three colors (red, black and blue) and we use dark shades to encode a 0 and light shades to encode a 1. Thus, a binary sequence can be sequentially embedded into the cover text. Notice that the embedding rate is comparatively higher than the rate of inter-line or inter-word space modulation methods. Furthermore, according to the desired level of robustness against digital-to-analog-to-digital (D-A-D) conversion, e.g. the P-S channel, one can choose which characters will be used to embed the data and which ones will be ignored. Indeed, small characters, like periods and commas, may not be good information carriers for printed text documents. This problem is less likely to occur in an electronic-only environment.

\[
\text{VAMOS A TRABAJAR} \quad 010110 \quad 01000101 \\
\text{(a)} \quad \text{VAMOS A TRABAJAR} \quad \text{(b)}
\]

Figure 4.3: Color Index Modulation: (a) original text; (b) marked text (exaggerated).

Obviously, this method satisfies requirements \textbf{R1}, \textbf{R2}, and \textbf{R3}. Requirement \textbf{R4} is also satisfied because we know from the HVS characteristics that slight luminance variations will not be noticed by the human eye. Moreover, luminance variations over bright (and dark) backgrounds are less visible than luminance variations over gray backgrounds [64]. Luckily, most text documents are written using dark color characters over a bright background. By using a modern word processor, one can easily verify that in an electronic-only environment this method can embed up to 4 bits per character (using color shades from 0 to 15) while still satisfying requirement \textbf{R4}. If the electronic stego text document is to be printed and scanned, we expect an embedding rate of 1-2 bits per character. Concerning the automated processing of printed and scanned stego text documents, one can use one of the numerous existing document segmentation algorithms depending on the targeted application and the a priori knowledge of the document layout [65]. The correct segmentation of individual characters is essential in order to estimate their color; however, OCR is not necessary. Therefore, our method also satisfies requirements \textbf{R5} and \textbf{R6}.

From the point of view of the practical applications, this method is efficient for authentication and tamper proofing because of the following reasons:

- As indicated above, robustness against D-A-D conversion can be accomplished by selecting the set of characters to be used to embed the data. Further improvement can be obtained by grouping small sequences of characters or by using error control codes.
• Repetitive embedding can be used to recover the hidden data from incomplete text documents and to perform both synchronization and channel compensation prior to decoding.

The reader is referred to Deguillaume et al. [62, 63] for further information on the last two points.

**Two-level quantizer.** The easiest method to embed color information is by using a two-level quantizer. In this approach, we fix a reference color representing bit 0. A good choice is to use the document’s original color (most of the time black). We choose a lighter shade in order to represent bit 1. We show in Figure 4.4 a concrete example of this method, where we used \( Q_0(x) = 0 \) (black) and \( Q_1(x) = 46 \) to mark the characters. The experimental performance evaluation of this method, for both electronic and printed text documents, is given in Section 4.6.1.

![Two-level color index modulation](image)

**Multilevel quantizer.** The above method can be easily extended to multiple levels. Instead of using two color levels, one may use four or eight levels, i.e. a multilevel quantizer. The data embedding rate could in this case be multiplied by a factor of two or three, according to the number of embedded bits per character. Naturally, for practical applications, this method relies on the quality of the printing and scanning machines. With the advent of very high quality printers, e.g. very high resolution ink jet printers, and very high dynamic range scanners, this extension should not be disregarded.

### 4.4.2 Location Index Modulation (LIM)

In this method, the stego text is obtained via (4.1), for \( \alpha' = 1 \) and defining the character feature \( X \) to be location. The spatial location of each character is defined with respect to a two-dimensional (2-D) orthogonal coordinate system, i.e. \( X = (X^h, X^v) \). Here, we assume that the same coordinate system can be used for locating the original character, the stego character, and the noisy stego character.
In its simplest form, this method quantizes either the horizontal coordinate $X^h$ (also known as inter-character space modulation [66]) or the vertical coordinate $X^v$. However, it is also possible to quantize both coordinates at the same time. In this case, one should think of $Q_m(\cdot)$ in (4.1) as a 2-D vector quantizer acting on both horizontal and vertical coordinates. We show in Figure 4.5 some examples of quantizers for this scheme.

![Figure 4.5: Location Index Modulation ($N = 1$): (a) original character $x = A = (x^h, x^v)$; (b) marked character $Q_1(x)$ by horizontal shifting; (c) marked character $Q_1(x)$ by vertical shifting; (d) marked character $Q_0(x)$ by combined horizontal and vertical shifting.](image)

It is possible to generalize this method to consider groups of characters, such as words or lines of text, instead of individual characters. Such a generalization is based on (4.2), where we take $\alpha' = 1$ and the feature vector $X^N$ is obtained by concatenation of the individual component character locations $X_n = (X^h_n, X^v_n)$, $n = 1, 2, \ldots, N$. In this case, one has many degrees of freedom for designing the vector quantizer $Q_m^N(\cdot)$. For example, if $X^N$ represents a word, then one could design $Q_m^N(\cdot)$ in such a way that all characters of $X^N$ are shifted horizontally by the same amount and in the same direction. This technique is also known as word-shift coding [56]. Similarly, another well-known technique, namely line-shift coding [56], is a special case of the described vector quantization method, where $X^N$ represents a line of text and $Q_m^N(\cdot)$ is such that all characters of $X^N$ are shifted vertically by the same amount and in the same direction. We show in Figure 4.6 a 2-D representation of possible quantizers for this vector quantization scheme.

### 4.4.3 Halftone Index Modulation (HIM)

This method, which is similar to the one proposed by Matsui and Tanaka [67], relies on halftoning. Here, we restrict our discussion to black and white printers, i.e. grayscale printing.

In order to simulate a given gray shade a halftone printer uses a halftone screen. Our method exploits the fact that there exist several possible choices for the halftone screen leading to the same gray shade. Therefore, one can use this property to hide data on each text character by using a different halftone screen according to the message $m$ that one

---

2An exact graphical representation is not possible since one would need to represent $N$-dimensional vectors $X^N = (X_1, X_2, \ldots, X_N)$, where each component is a complex number (the real part representing the horizontal coordinate and the imaginary part representing the vertical coordinate).
wishes to embed. Typical halftone screen characteristics that can be exploited for this purpose are: screen angle and screen dot shape (elliptical, round, square).

Again, the stego text is obtained via (4.1) for $\alpha' = 1$. However, the quantizer $Q_m(\cdot)$ must be seen as a vector quantizer acting on the gray values of the pixels making up each character. We show in Figure 4.7 an example of this method where a screen angle of 0° is used to encode bit 0 whereas a screen angle of 45° is used to encode bit 1. The major strength of this method is that all characters in the stego text will have the same grade shade. On the other hand, unless combined with the color index modulation method, this method is intended mainly for printed documents. For example, if $|M| = 2$, one may use the set of grade shades $\{Q_0(x) = 45, Q_1(x) = 46\}$ to embed binary data in the electronic version of the text document, and a halftone pattern screen together with two screen angles to embed binary data into the printed version of the text document.

4.4.4 Hybrid Schemes

A natural extension of the schemes described by (4.1) and (4.2) is to consider simultaneously multiple character features instead of a single one. In fact, this idea was already introduced in the previous Subsection while describing LIM. Indeed, one can consider the horizontal coordinate of a character as its first feature and the vertical coordinate as its second feature. Notice that these two features are independent from each other. Clearly,
the main advantage of combining multiple independent features is the higher data embedding rate of the resulting scheme. For example, the schemes in Figures 4.5b and 4.5c have a maximum data embedding rate of 1 bit/character, whereas the scheme in Figure 4.5d has a maximum data embedding rate of 2 bits/character. Another example of a (true) hybrid scheme with even higher data embedding rate (maximum 3 bits/character) is the one that combines the CIM scheme shown in Figure 4.3 with the LIM scheme shown in Figure 4.5d.

Hybrid schemes may also have other advantages due to the fact that different features have different properties. For example, it is known that CIM is less robust to photocopying than LIM; therefore, one can build a hybrid scheme capable of authenticating the contents of a text document (based on LIM and robust text hashing), and capable of discerning the original text document from its copies (based on CIM).

4.4.5 Error Control Coding for Print-Scan Channels

The quantization-based methods explained above, which are by themselves channel codes, may not be completely robust to one-time or many-times printing and scanning, i.e. the decoder’s output may contain some errors. In fact, it can be verified by experimentation that there is a trade-off between the invisibility of the watermark and the decoding accuracy. Typical examples of this problem relate to the use of printers, scanners, copy machines, fax machines, etc.

In order to reduce the error rate to an acceptable level, an outer layer of coding can be used. The correct design of such an outer layer takes into account the equivalent channel formed by the quantization encoder, the concatenation of the employed P-S channels, and the quantization decoder. Moreover, the operation of the overall decoding machinery may require some modifications in order to get full benefit of soft-decision decoding techniques. For example, the quantization decoder may be modified so that it outputs soft estimates rather than hard estimates. Notice that the equivalent channel is, in general, different according to the used quantization technique (e.g. CIM, LIM, etc.). Accurate modeling of this channel for a particular quantization technique is challenging and there exist only few works [29, 68] tackling this problem.

For instance, if CIM and halftone printing are used, one may model the equivalent channel as follows (Figure 4.8). The quantized signals are mapped from the intensity space to the halftone space using a halftone encoder:

\[ f_{HT} : \mathcal{Y}^N \rightarrow \tilde{\mathcal{Y}}^{A \times N}, \]

where each character symbol from \( \mathcal{Y} \) is reproduced by a halftone pattern from \( \tilde{\mathcal{Y}}^A \) of size
$a_1 \times a_2$ with $A = a_1 \cdot a_2$. The halftone space is defined to consist of black and white dots $\hat{Y}^A = \{0, 255\}^A$ that are reproduced by a printing device with resolution $r_p$. We assume that a scanner performs a mapping of this pattern into a new one with resolution $r_s$ with a given size $a'_1 \times a'_2$ and $A' = a'_1 \cdot a'_2$. Two possibilities exist depending on the relationship $r_s \geq r_p$.

![Figure 4.8: Channel model for electronic-printed text document authentication.](image)

between the printing and scanning resolutions $r_p$ and $r_s$. If $r_s = r_p$, the so-called indirect decoding is applied where the halftone pattern $\hat{V}^{A' \times N}$ is first mapped to the intensity $V^N$ using an estimator of intensity (halftone decoder known also as inverse halftone decoder):

$$g_{HT} : \hat{V}^{A' \times N} \rightarrow V^N$$

and then decoded. In this case, $A = A'$ and the P-S channel is modeled as a parallel binary symmetric channel with a given transition probability. The drawback of this decoding approach is related to the corresponding consequences of the data-processing inequality. If $r_s > r_p$, one can perform direct decoding using a halftone pattern codebook thus avoiding the mapping to the intensity space.

It should also be pointed out that the channel model developed in Section 2.3.3, in the context of multilevel 2-D bar codes, can be used to accurately model the P-S channel between $Y^N$ and $V^N$ when CIM is used. The main idea here is to consider a character symbol as a 2-D bar code symbol. The resulting channel model is:

$$V = \rho(Y) + Z,$$

where $\rho : \mathcal{Y} \rightarrow \mathbb{R}$ is a non linear function and $Z$ represents zero-mean additive generalized Gaussian noise whose parameters are determined by the input $Y$, i.e. $(Z|Y = y) \sim \mathcal{GGD}(0, \sigma_{Z|Y}(y), \gamma_{Z|Y}(y))$, where $\sigma_{Z|Y}(y)$ and $\gamma_{Z|Y}(y)$ are the noise standard deviation and shape parameter given $Y = y$. In particular, as was shown in Chapter 2, if a multilevel quantizer is used, then a multilevel coding (MLC) together with a multistage decoding (MSD) can be designed to reliably communicate information through any P-S channel.

Finally, notice that the use of an error control coding (ECC) scheme together with a text data hiding scheme decreases the data embedding rate of the overall scheme. For example, if the text data hiding rate is 2 bits/character and the rate of the ECC scheme is $1/2$, then the effective data embedding rate is 1 bit/character.
4.5 Robust Hashing of Text Documents

A robust text hashing function $\theta$ takes as input a text object $X^N$ and a secret key $K_H$ to give the hash $B^S = \theta(X^N, K_H)$. The text object could be either a character, a word, a sentence, a paragraph, a line of text, a text fragment, or even the whole text document. The hash $B^S$ is required to be invariant under unintentional/intentional legitimate modifications of the text document such as conversion between electronic formats, data hiding, and typical handling operations that include printing, scanning, photocopying, faxing, etc.

We will consider two types of text hashing techniques. One attractive feature of these techniques is that they are compatible with character feature text data hiding methods such as CIM and LIM.

4.5.1 OCR + MAC Text Hashing

This text hashing technique is based on OCR and a classical cryptographic message authentication code (MAC). As shown in Figure 4.9, the main idea is to apply OCR to the text document in order to obtain its ASCII representation; and then, using the secret key $K_H$, to compute the MAC of this representation in order to obtain the desired hash. As will be shown in Section 4.6.2, the use of OCR provides, in general, good robustness against legitimate modifications. However, since this technique highly relies on the accuracy of the employed OCR tool, it completely fails when the OCR tool makes a mistake (this is because classical cryptographic MACs for similar inputs are generally quite dissimilar). This is, however, not a very important issue since it is possible to tune the OCR engine to be more robust against the considered class of legitimate modifications.

![Figure 4.9: OCR + MAC text hashing.](image)

4.5.2 Random Tiling Text Hashing

Inspired by original work for images [69], we describe in the following paragraphs a new robust text hashing algorithm. Let $x^N = (x_1, x_2, \ldots, x_N)$ represent an input text object and $k_H$ be a secret key. We suppose that $x^N$ comes either from an electronic support (vector graphics representation of a text document) or from a scanned image of a printed text document. In our algorithm, we use $k_H$ as a seed for the pseudorandom number
generator used in all steps requiring random quantities. The following algorithm produces the hash for a single text object.

1. Preprocess the image containing the text object so as to correct a possible skew.

2. Segment and convert the image into a bitmap composed of only black and white pixels. We suppose that the text object \( x^N \) belongs to a well-defined region \( \mathcal{R} \) in this image. For simplicity, we fix the shape of \( \mathcal{R} \) to be a rectangle (see Figure 4.10).

3. Generate at random \( P \) rectangles \( \mathcal{R}_p = \{(i, j) : 1 \leq i \leq I_p, 1 \leq j \leq J_p\} \), where \( I_p \) and \( J_p \) are, respectively, the height and width in pixels of the \( p \)-th rectangle \( \mathcal{R}_p \), \( p = 1, 2, \ldots, P \). We assume that each rectangle \( \mathcal{R}_p \) is randomly positioned inside \( \mathcal{R} \), i.e. \( \mathcal{R}_p \subset \mathcal{R} \), and that \( \bigcup_{p=1}^{P} \mathcal{R}_p = \mathcal{R} \). In Figure 4.10, we schematize the generated random rectangles for two kinds of text objects.

4. Compute
\[
\mu_p = \frac{1}{\sum_{r \in \mathcal{R}_p} \lambda_{k_H}(r)} \sum_{r \in \mathcal{R}_p} \lambda_{k_H}(r) l(r),
\]
for \( p = 1, 2, \ldots, P \), where \( l(r) \) is the luminance value (0 or 1) of the pixel located at \( r = (r^h, r^v) \) and \( \lambda_{k_H}(r) \in \{0, 1\} \) is a key-dependent weight for the same pixel. If \( \lambda_{k_H}(r) = 1 \) for all \( r \in \mathcal{R}_p \), then \( \mu_p \) is simply the sample mean.

5. Compute the intermediate hash \( \tilde{b}^P = (\tilde{b}_1, \tilde{b}_2, \ldots, \tilde{b}_P) \) by randomly picking \( P \) thresholds \( \tau_p(k_H) \in [0, 1], p = 1, 2, \ldots, P \) and defining \( \tilde{b}_p \) as:
\[
\tilde{b}_p = \begin{cases} 
0 & \text{if } \mu_p < \tau_p(k_H), \\
1 & \text{otherwise}.
\end{cases}
\]

6. Produce the final hash \( b^S = \theta(x^N, k_H) \) by randomly choosing an index set \( \{p_1, p_2, \ldots, p_S\} \subset \{1, 2, \ldots, S\}, S \leq P \), and letting \( b^S = (\tilde{b}_{p_1}, \ldots, \tilde{b}_{p_S}) \).

Figure 4.10: Random tiling text hashing: (a) character-based, \( x = W \); (b) word-based, \( x^7(\text{H,a,s,h,i,n,g}) \).
4.6 Experimental Results

4.6.1 Text Data Hiding

In this part, we describe a practical implementation of CIM. As explained in Section 4.4.1, this method can be used for both electronic and printed text documents.

The implementation of this method in an electronic-only environment is straightforward. In our experiments, we implemented a prototype for Microsoft Word documents capable of embedding and extracting any arbitrary message. Assuming perfect synchronization when reading the marked characters, our prototype was able to extract the embedded messages without any errors. Thus, for this case, the use of error control codes for the reliable extraction of an embedded message is not needed. We also verified that the conversion from the DOC format to the PDF or PS formats retains the color information of each character. Our implementation was also able to successfully extract the embedded message from the PDF or PS versions of a DOC document.

Now, we describe an extended implementation of color index modulation for text documents that are subject to D-A-D conversion. This implementation considers only a two-level quantizer, but it can be readily extended to consider multilevel quantizers. In Table 4.1, we list the exploited equipment for performing the experiments. The default printer parameters (resolution, screen frequency, halftone algorithm) were used for printing the electronic text documents. For scanning the printed text documents, we used a resolution of \( r_s = 600 \) ppi, grayscale mode, 8 bits of bit-depth, full dynamic range (from 0 to 255 according to the bit-depth setting), \( \gamma \)-correction set to 1, and an unsharp mask filter of high level according to each scanner’s driver interface.

For the sake of simplicity, we first selected a set of random black electronic texts written using the Latin alphabet (A, B, ..., Z, a, b, ..., z), common punctuation and special symbols (comma, period, colon, semicolon, -, ?, !, “”, ’, ( ), <, >, @, |), numbers (0, 1, ..., 9), and arithmetic operators (+, −, *, /, =). Our practical system can nonetheless work with other languages using other alphabets. We used Arial font characters of size 10 pt (1 pt \( \approx 1/72 \) in). In order to be robust against printing and scanning, some characters were deliberately not used for embedding information. These characters are the following: comma, period, colon, semicolon, “”, ’, ’, −. Secondly, an equal number of arbitrary messages were embedded into the electronic texts using the two-level color index modulation. The marked electronic texts were subsequently printed and scanned with the equipment listed in Table 4.1. Finally, the scanned electronic texts were processed in order to extract the embedded messages.

<table>
<thead>
<tr>
<th>Model</th>
<th>Type</th>
</tr>
</thead>
<tbody>
<tr>
<td>HP Color LaserJet 4600</td>
<td>laser printer</td>
</tr>
<tr>
<td>Epson Perfection 3170 Photo</td>
<td>CCD scanner</td>
</tr>
<tr>
<td>Epson Perfection 4990 Photo</td>
<td>CCD scanner</td>
</tr>
<tr>
<td>Canon LiDE 50</td>
<td>CCD scanner</td>
</tr>
</tbody>
</table>

Table 4.1: Equipment used for experimentation.
The extraction process can be divided in three parts: segmentation of characters, demodulation of the character feature (in this case, color), and quantization-based decoding.

For the segmentation of characters, we assumed good compensation (document deskewing) of a possible document misalignment while scanning. The implemented segmentation algorithm is based on off-the-shelf methods for character segmentation [56, 70]. We briefly explain the involved steps (see also Figure 4.11):

i. Apply a $\gamma$-correction factor of 2.6,

ii. Identify the text line boundaries by computing the luminance vertical profile (i.e. by projecting the pixel luminance values onto the Y-axis),

iii. Provisionally eliminate the halftone patterns via a local median filter,

iv. Identify the text character boundaries on each line using the SUSAN edge detector [70],

v. Identify the areas where the demodulation of features will be performed.

![Figure 4.11: Segmentation of characters.](image)

Due to the used printing technology (halftoning), one has several options for the demodulation of the character feature, e.g. computation of the character’s average luminance or analysis of the character’s halftone pattern (in this case, quantify whether a halftone pattern is present or not). We tested both approaches and found that the latter is more robust against printing and scanning.

Finally, the two-level quantization decoder was implemented by experimental optimization of a threshold.

The obtained results were similar for all the exploited scanners. For this reason, we only show in Table 4.2 the results for the Epson Perfection 3170 Photo scanner. These results were obtained using halftone pattern demodulation for texts of $J = 4104$ characters. In this table, $Q_0(x)$ and $Q_1(x)$ represent the luminance values employed to mark the characters.

### 4.6.2 Robust Text Hashing

In this part we present the experimental results for the two text hashing methods described in Section 4.5. We used standard office equipment (printer, scanner, copy machine, fax machine) to perform our tests. All of the electronic images containing text objects were created/processed at 600 ppi.
Table 4.2: Performance of two-level color index modulation ($J = 4104$).

<table>
<thead>
<tr>
<th>$Q_0(x)$</th>
<th>$Q_1(x)$</th>
<th>Error count</th>
<th>Error rate</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>41</td>
<td>1342</td>
<td>32.7%</td>
</tr>
<tr>
<td>0</td>
<td>46</td>
<td>824</td>
<td>20.1%</td>
</tr>
<tr>
<td>0</td>
<td>51</td>
<td>315</td>
<td>7.7%</td>
</tr>
<tr>
<td>0</td>
<td>56</td>
<td>120</td>
<td>2.9%</td>
</tr>
<tr>
<td>0</td>
<td>61</td>
<td>62</td>
<td>1.5%</td>
</tr>
<tr>
<td>0</td>
<td>66</td>
<td>23</td>
<td>0.6%</td>
</tr>
<tr>
<td>0</td>
<td>71</td>
<td>9</td>
<td>0.2%</td>
</tr>
<tr>
<td>0</td>
<td>76</td>
<td>0</td>
<td>0.0%</td>
</tr>
</tbody>
</table>

For the implementation of OCR + MAC text hashing we used ABBYY FineReader as OCR tool and HMAC SHA-1 truncated to 80 bits as MAC. Assuming that a text line of 80 characters can reliably store 80 hash bits (see Section 4.4.5), this implementation takes into account the rate requirement of the text data hiding part of the self-authentication system, namely that $R_H \leq R_{DH}$.

The following parameters were used for the implementation of random tiling text hashing: $P = 1024$, $\lambda_{kn}(r) = 1$ for all $r \in \mathcal{R}_p$, and $S = P$ (in fact Step 6 in Section 4.5.2 is replaced by $b^S = b^P$). The width and height (in pixels) of a random rectangle $\mathcal{R}_p$ were drawn uniformly at random from the intervals $[5, 10] \cap \mathbb{Z}$, $[5, 50] \cap \mathbb{Z}$, respectively. We do not claim, however, any optimality of the used parameters (e.g. the length of the hash is still relatively long in the current implementation).

The considered text objects were text lines (Arial font, 10 pt). A sample of five text objects is shown in Figure 4.12.

![Figure 4.12: Sample text lines.](image)

Two classes of modifications were considered in the scope of this work: legitimate and illegitimate modifications. The legitimate modifications include electronic format conversion (DOC ↔ PS ↔ PDF), printing and scanning, photocopying, and faxing. The illegitimate modifications (see Figure 4.13) include tampering of text lines by adding one new character, by suppressing one character, by replacing one character with a visually different one, and by replacing one character with a visually similar one.

Concerning the legitimate distortions, we show in Figures 4.14 and 4.15, the obtained results for OCR + MAC text hashing and random tiling text hashing, respectively. These figures should be interpreted as follows. Rows (or columns) from 1 to 5 represent the sample text lines of Figure 4.12. Rows (or columns) from 6 to 10, in this order, correspond also to the sample text lines 1 to 5 of Figure 4.12 but after a legitimate distortion has been
A text document authentication system aims at deciding whether a given text document is authentic or not. The decision about authenticity is performed at the global level, meaning that the system gives only a binary decision about the entire document: authentic or fake. On the contrary, if a system makes decisions at the local level, we refer to it as a document tamper-proofing system. Thus, text document authentication and tamper-proofing

6
7
8
9
10
(a)
(b)
(c)
(d)

Figure 4.13: Sample of forged text lines: (a) by adding one new character; (c) by suppressing one character; (c) by replacing one character with a visually different one; (d) by replacing one character with a visually similar one.

We use the normalized Hamming distance $d_H(b_1^S, b_2^S)$ to compare any two hash values $b_1^S$ and $b_2^S$. Recall that the desired properties for a robust text hashing method should be the following:

i. $d_H(b_1^S, b_2^S) \approx 0$ for any two similar text objects whose hashes are $b_1^S$ and $b_2^S$,

ii. $d_H(b_1^S, b_2^S) \approx 0.5$ for any two dissimilar text objects whose hashes are $b_1^S$ and $b_2^S$.

Figure 4.14: Normalized Hamming distances for legitimate distortions and OCR + MAC text hashing: (a) electronic format conversion; (b) printing and scanning; (c) photocopying; (d) faxing.

We observe from Figures 4.14 and 4.15 that both text hashing methods show good effectiveness for the tested legitimate distortions. However, from Figure 4.14d we observe
that when the OCR tool makes a mistake, e.g. a punctuation mistake, recognition of two spaces instead of one, etc., then the hash of a text line is completely different from the one obtained from a similar text line.

Concerning the illegitimate distortions, we show in Figures 4.16 and 4.17, the obtained results for OCR + MAC text hashing and random tiling text hashing, respectively. These figures should be interpreted as follows. Rows (or columns) from 1 to 5 represent the sample text lines of Figure 4.12. Rows (or columns) from 6 to 10 correspond to the forged text lines shown in Figure 4.13. We observe from Figures 4.16a, 4.16b, 4.16c and 4.17a, 4.17b, 4.17c that both text hashing methods show good effectiveness for the tested illegitimate distortions. From Figures 4.16d and 4.17d, we observe that only OCR + MAC text hashing is able to handle the corresponding illegitimate distortion. Random tiling text hashing completely fails in this case. However, notice again, from Figure 4.16d, that OCR + MAC text hashing can also fail if the OCR tool makes a mistake. In the case at hand, the OCR tool recognized the word authentication as authentication due to the employed spell-checker. For this reason it is highly recommended to turn off this feature and recognize text symbol by symbol.
4.7 Conclusions

In this Chapter, we studied the particular problem of text document authentication via robust visual text hashing and text data hiding.

Firstly, we addressed the problem of limited data embedding rate of current text data hiding methods by introducing a novel theoretical framework for studying such systems. We explained how the text data hiding problem can be seen as an instance of the well-known G-P problem. The main idea was to consider a text character as a data structure consisting of multiple quantifiable features such as shape, location, orientation, size, color, etc. The power of this framework was demonstrated by showing that previous text data hiding techniques, namely open space methods and character feature methods, are particular cases of a general quantization-based text data hiding technique, which is, in turn, a practical implementation of the G-P text data hiding scheme. We also presented CIM as a new method for semi-fragile data hiding of electronic and printed text documents. The experimental work confirmed that this method has high perceptual invisibility, high information embedding rate, and is fully automatable. In view of the obtained results it can be stated that color index modulation is suitable for many text document authentication and tamper proofing applications.

Secondly, we studied two text hashing algorithms, namely OCR + MAC text hashing and random tiling text hashing, that are particularly well suited for the considered problem. In particular, we showed by experimentation that OCR + MAC text hashing shows better applicability than random tiling text hashing. However, we have also observed that the OCR + MAC text hashing method highly relies on the accuracy of the OCR tool. Nonetheless, we believe that the OCR tool can be tuned in order to improve its robustness, e.g. by not using a spell-checker and ignoring multiple blank spaces. Moreover, the experimental work also confirms that both text hashing algorithms are robust against typical legitimate document distortions that include electronic format conversion, printing, scanning, photocopying, and faxing.

The main results of this Chapter have been published in [104, 105, 106].
5 Authentication of Biometric Identification Documents

5.1 Introduction

IDENTIFYING PEOPLE and protecting access or privileges to specific resources are among current needs of our modern society. Generally, an authority may grant certain rights to individuals, and provides these individuals with an identification document. When such a person desires to exercise these rights, there will be a challenger who, as the name suggests, defies the rights of the person. The person then presents the provided identification document to the challenger as a certificate that this person has the required rights. The challenger must then make a decision: either the presented identification document is valid, or it is not.

A problem is that technological advancements can be used by attackers for malicious purposes, including attempts to duplicate, alter or otherwise manipulate identification documents to circumvent the regular system of rights verification. In order to make a correct decision, the procedures and technical aspects of identification documents must be sufficiently advanced to prevent attackers from succeeding.

Current systems for authorization of persons are based on secure identification documents. Ways to secure such documents include the use of special inks [71], special paper [72], anti-copying visible patterns [73], embedded holograms and microtext [74]. Besides the high cost of an identification document that is protected with these methods, usually expensive equipment is required to perform the validation of protected identification documents; thus rendering the cost prohibitive for some applications, as well as limiting the possibility of verification to a restricted circle of authorized parties. Moreover, most of these techniques are based on proprietary undisclosed designs, which is not considered a good security policy, according to Kerckhoff’s principle.

Based on the obtained results in Chapters 3 and 4, we propose a novel low-cost approach to document security, and particularly authentication of identification documents, based on robust visual hashing, data hiding, biometrics and portable devices with optic equipment.

The rest of this Chapter is organized as follows. Section 5.2 presents an overview of the assumed identification framework. Section 5.3 states requirements for a modern
approach to identification based on biometrics and identification documents. In Section 5.4, we formally define the problem of interest and discuss two possible system designs. A practical scheme is proposed in Section 5.5 and in Section 5.6 we demonstrate by experimentation the feasibility of the proposed scheme. Finally, we draw conclusions and suggest future work in Section 5.7.

5.2 Identification Framework

5.2.1 Biometrics

Historically, person identification has been based on: 1) what a person possesses, such as a document or a card; 2) what a person knows, for example, a password or a personal identification number; or 3) biometrics, inherent physiological or behavioral characteristics. A classical identification document belongs to the first category. However, possessing a card is usually not sufficient to guaranty that a given person is its legitimate owner, since cards can always be stolen and misused. This is why identification documents usually contain information that links them to their legitimate holder – usually a photograph, and possibly other biometric data such as a handwritten signature or a fingerprint.

In general, typical biometrics that can be used are: face, fingerprint, hand geometry, iris, retinal scan, voice, handwritten signature or DNA [75]. Biometric identification is based on the unique nature and extreme richness of the human features enumerated above. Photographs present on passports or identification documents (IDs) can be seen as physiological biometrics, as well as fingerprints, which have been used by the police for criminal investigation since the late 19th century.

Biometrics-based person identification is believed to achieve the highest security and reliability. It appears today that biometrics could be used for person identification and access control in banking, border control, loyalty programs, time and attendance monitoring, etc. During the last few years significant efforts have been devoted to the theoretical investigation as well as to the development of practical methods for identification document security and biometrics-based person identification [76, 77, 78].

5.2.2 Personal Data

We refer to personal data as any information in an identification document that is not a representation of biometric data, but rather any information in text form. The personal data can typically include a name, position, date of birth, contact information, etc.; and one or more identifiers, such as any number, code or time-stamp. Such an identifier may be used to uniquely identify the document itself, where and when it was issued, the authority which issued it to the specified person; or to provide a link to other sources of information associated with the ID holder. The challenger might use this external information about either the identification document or the person in the decision process. Typically, personal data are printed in clear form on the document. More recently, personal data are additionally stored in an encrypted form for machine-based inspection.
5.2.3 Document-Based Identification

A complete setup of the assumed framework for authorization of people is depicted in Figure 5.1. According to the assumed framework, the role of an identification document is to prove that a given person has been authorized by an authority. As such, the identification document is of critical importance to the proper functioning of the system.

For the challenger, the mere existence of an identification document should not be sufficient to consider it a proof that the person holding it is indeed authorized. As explained in Section 5.2.1, the document needs to be linked to the intended holder. Using biometrics, it is assumed that this task can be performed up to sufficient accuracy. In our assumed setup, the biometric data is present on the ID and the challenger is expected to verify that this data is indeed linked to the person to be identified. This task could be performed with a mobile device with biometric verification capabilities similar to those of the SecurePhone [79]; the biometric data would be on the ID instead of on a SIM card.

Once the challenger has established that the document is indeed related to the person, a decision could be made using the available information, as detailed in Section 5.2.2, regarding the authenticity of the document and the authorization of the person. We will not go into the details of the authorization here, but assume that this can successfully be handled once the identification document has been authenticated. The open issue to address is then to find a convenient document authentication framework that the challenger can use to verify that the presented document is authentic, i.e. it is truly an original document provided by the authority.

Even though a document might be original (provided by the authority), this does not mean that it is unmodified. We consider legitimate distortions as modifications such as normal wear and tear, scratching, and operations carried out as part of the normal handling such as printing and scanning. Legitimate distortions introduced by normal handling are expected to preserve a sufficient degree of quality of the document to carry out verification.

![Figure 5.1: The relations between an authority, a person, a challenger and an identification document.](image-url)
procedures; i.e. a few scratches are acceptable, but scratching out an entire image is clearly not.

An illegitimate document is a document that is not at all provided by the authority, or a document that has illegitimate distortions, by which we mean that it has undergone a modification that is not legitimate. This would be any form of modifications whereby insufficient document quality is preserved, leading the challenger to the negative decision regarding the document’s authenticity. For example, a possible attack scenario would be to replace or modify the personal or biometric data on an identification document. Because in habitual identification documents a cross-link between the personal and the biometric data is absent, the probability of making an incorrect decision is increased. We propose to add such a protective link using robust visual hashing and data hiding technologies in order to increase the accuracy of modern person identification systems.

5.3 Requirements and Proposed Approach

5.3.1 Mobile Device Architecture

The use of digital technologies, networks and computers, and the increasing dynamics and mobility in the modern world recently revealed unprecedented threats leading to an urgent need for securing identification documents. Among the major security issues the most important are: the ease with which exact copies of digital content can be produced without authorization; the effectiveness with which high quality counterfeit identification documents can be made with common document editing tools; and the ease with which the true originator or owner of a document can be faked for fraudulent purposes.

Despite the increasing risk of fraud, identification documents still remain the most widely used means of person identification. Thus, there is a great need for systems that allow accurate and quick document authentication.

In this work, we propose a unified system architecture dealing with the protection of identification documents and allowing to establish their authenticity. The proposed system architecture is shown in Figure 5.2. The identification document is acquired using the sensors of a portable device, e.g., a digital camera. The resulting image is sent to an authentication server, which either directly establishes the document’s authenticity, or compares the sent data with those stored in a database in order to reach a decision. The verification result is sent back to the portable device and displayed for the end-user.

5.3.2 Identity Document Design: Link Between Biometrics and Personal Data

As discussed in Section 5.2.3, the identification document should create a link between the biometric data and the personal data. One can consider three practical variants of identification documents with this security feature: 1) a remote database as in Figure 5.3a, 2) an on-document storage module (SM) as in Figure 5.3b or 3) an on-document cross-storage based on robust visual hashing and data hiding, as in Figure 5.3c.

In the first system, the system performs a comparison of the biometric and personal data with those stored in a trusted remote database. This approach requires keeping a
database and a real-time connection during the identity verification. Keeping a remote database can be impractical and raises privacy issues in many common scenarios as gathering information about people is restricted under current laws in most of the western world [80].

The two remaining variants consist in the storage of the link between the biometrics and the personal data directly on the identification documents. In this case, the link should be stored in a way that is resistant against forgery. The second system is based on a secure storage module, that can be an electronic chip (smart cards), a two-dimensional (2-D) bar code printed or laser engraved on the document’s surface, readable optical storage modules and more recently RFID tags.

The third system does not need any additional physical modules or devices to link biometrics to personal data. It uses the authentication techniques presented in Chapters 3 and 4 that can create a reliable cross-link between biometrics and personal data.

One of the main advantages of the last two systems is the possibility to adapt them to most of the existing classical identification frameworks based on identification documents or smart cards. Additionally, the third system does not require a change of the document layout, which is an important factor for standardization. Another advantage of storing biometrics and the link to personal data on the document itself is the possibility to perform fast and convenient identity verification without the need for a direct link with a centralized database. Moreover, better privacy could be achieved by holding all the pri-
vate information on the document itself. For example, since October 2004, the United States Secretary of State requires machine readable passports for travelers entering the country without a visa under the Visa Waiver Program (VWP) [81]; the Departments of Homeland Security and of State requested an extension consisting of the inclusion of biometric features in passports issued by countries participants of the VWP. Identification documents based on on-document storage of biometrics to make them less vulnerable with respect to illegal use or falsification are probably the main trend to be followed by most governments.

Furthermore, a multimodal biometric system can benefit from data hiding techniques to achieve cross-modal information embedding, meaning that the information from one mode can be embedded into one or several other modes [82, 83]. By making multiple links between the different modes, such cross-modality renders the system even more difficult to break. The key idea of our identification framework is then to generate an identification document, which adds a link between the biometric and personal data. This link is based on data hiding in the different modalities, where each piece of information stores extra data about its complement. Taking into account the limited embedding rate compared to the amount of information to embed, we propose not to embed the entire data of one modality into the other, but either a robust hash of it.

One advantage of such a design is related to the document manufacturing and personalization that can be combined in one technological step. This facilitates document production and leads to cost reductions. Note that some of the elements from the two categories above are physically printed on the document in clear form for visual inspection. The printed elements could be textual personal data, the document’s identification number, as well as a photograph of the face and the handwritten signature. The backward compatibility of the proposed approach with classical identification documents relies on these printed data, which we will refer to as “human-readable data”. On the opposite, for automatic document verification all the data elements listed above need to be encoded in some way and embedded in the document using data hiding. The watermark could be embedded either on the whole document area (including its background), into the printed photograph and in text data. The retained data hiding technique depends on its maximal embedding rate with respect to the amount of information to embed, as well as on the desired level of robustness.

The authentication relies on the encryption of the hidden data in an asymmetrical manner with the private key of the authority issuing the document. Based on the printed elements and embedded hidden data, the system performs the authentication of a document. The authentication of an identification document should not rely on proprietary devices, which are difficult to standardize and distribute world wide. Recent trends in identification documents design consider publicly available devices such as mobile phones and personal digital assistants (PDAs). That is why one should match the security and storage requirements of identification protocols and data hiding technologies with the imaging facilities of portable devices. Development of such protocols and data hiding technologies is a great challenge and a subject of intensive research in the cryptographic and watermarking community [84].

Due to the desire to use portable devices such as mobile phones in the architecture, cross-storage based on robust hashing and data hiding is an attractive solution and will be explored in further detail.
5.4 Problem Formulation and Solution Approaches

We believe that despite the apparent variety of existing person identification systems based on biometrics, data hiding, and robust visual hashing technologies, only a few powerful basic principles can be used to theoretically analyze and guide the design of these systems. The task of identifying, introducing and applying these fundamental principles is a great challenge. The goal of the present Section is to apply the theoretical framework developed in Chapter 3 to the problem of authentication of biometric identification documents.

5.4.1 Authentication of Biometric Identification Documents

We consider the problem of authenticating a document that holds biometric and personal data within the framework developed in Section 3.2. The generic block diagram of an authentication system meant to solve this problem is shown in Figure 5.4.

![Figure 5.4: Block diagram of a generic biometric document authentication system.](image)

The authentication authority or encoder \( \varphi \) has access to the uniquely assigned secret key \( K \) that is uniformly distributed over the set \( \mathcal{K} = \{1, 2, \ldots, |\mathcal{K}|\} \), the non-causal realization of biometric data \( X_1^{N_1} \in \mathcal{X}_1^{N_1} \), and the personal data \( X_2^{N_2} \in \mathcal{X}_2^{N_2} \). As explained in Section 5.2.1, \( X_1^{N_1} \) can be any personal biometric. On the other hand, as detailed in Section 5.2.2, \( X_2^{N_2} \) corresponds to any personal information that can be represented in text form. The key \( K \), the biometric data \( X_1^{N_1} \), and the personal data \( X_2^{N_2} \) are used at the encoder \( \varphi \) to generate the tamper resistant data \( Y_1^{N_1} \in \mathcal{Y}_1^{N_1} \) and \( Y_2^{N_2} \in \mathcal{Y}_2^{N_2} \). The tamper resistant data \( (Y_1^{N_1}, Y_2^{N_2}) \) are communicated through a channel, which introduces some legitimate distortions described by two transition probabilities \( p(v_1^{N_1}|y_1^{N_1}) \) and \( p(v_2^{N_2}|y_2^{N_2}) \). The resulting distorted data are denoted \( V_1^{N_1} \in \mathcal{V}_1^{N_1} \) and \( V_2^{N_2} \in \mathcal{V}_2^{N_2} \). The challenger or decoder \( \xi \) makes a decision about the authenticity of the biometric identification document using \( (V_1^{N_1}, V_2^{N_2}) \) and \( K \). Thus, the document authentication system consists of the set \( \{\mathcal{X}_1^{N_1}, \mathcal{X}_2^{N_2}, \mathcal{K}, \mathcal{Y}_1^{N_1}, \mathcal{Y}_2^{N_2}, p(v_1^{N_1}|y_1^{N_1}), p(v_2^{N_2}|y_2^{N_2}), \mathcal{V}_1^{N_1}, \mathcal{V}_2^{N_2}\} \), the encoder’s allowable distortions \( D^{E_1} \) between \( X_1^{N_1} \) and \( Y_1^{N_1} \), the channel’s allowable distortions \( D^{C_1} \) between \( Y_1^{N_1} \) and \( V_1^{N_1} \), and the encoder-decoder pair \( \varphi : \mathcal{X}_1^{N_1} \times \mathcal{X}_2^{N_2} \times \mathcal{K} \rightarrow \mathcal{Y}_1^{N_1} \times \mathcal{Y}_2^{N_2} \), \( \xi : \mathcal{V}_1^{N_1} \times \mathcal{V}_2^{N_2} \times \mathcal{K} \rightarrow \{0, 1\} \), where \( i = 1, 2 \).

We formulate the problem of deciding whether the received data \( (V_1^{N_1}, V_2^{N_2}) \) are authentic or not as a binary hypothesis testing problem. One can assume that \( H_0 \) corresponds
to the hypothesis that the received data are authentic, and $H_1$ to the alternative hypothesis. Thus, the task of the authentication system is to decide which of the two hypotheses should be accepted given the realizations $(v_1^{N_1}, v_2^{N_2})$ and $k$. The hypothesis test problem can be stated as:

$$
\begin{align*}
H_0 : (V_1^{N_1}, V_2^{N_2}|K = k) &\sim P_{V_1^{N_1}, V_2^{N_2}|K}(v_1^{N_1}, v_2^{N_2}|k), \\
H_1 : (V_1^{N_1}, V_2^{N_2}|K = k) &\sim P_{V_1^{N_1}, V_2^{N_2}|K}(v_1^{N_1}, v_2^{N_2}|k), 
\end{align*}
$$

where $H_0$ corresponds to the decision 0 and $H_1$ to 1.

Similarly to Chapter 3, by assuming the use of the Neyman-Pearson test, one can obtain the complete system performance limits in terms of error exponents:

$$
\begin{align*}
P_M &\equiv 2^{-D\left(\frac{p^0_{V_1^{N_1}, V_2^{N_2}|K}}{p^1_{V_1^{N_1}, V_2^{N_2}|K}}\right)} \text{ for a fixed } P_F, \\
P_F &\equiv 2^{-D\left(\frac{p^1_{V_1^{N_1}, V_2^{N_2}|K}}{p^0_{V_1^{N_1}, V_2^{N_2}|K}}\right)} \text{ for a fixed } P_M.
\end{align*}
$$

### 5.4.2 Authentication Using Cross-Storage of Hashes

In this section, we consider all the elements of an authentication system based on the separation of robust visual hashing and data hiding. The considered identification document uses cross-storage of hashes as described in Section 5.3.2.

Let us again assume $i = 1, 2$. Referring to Figures 5.5 and 5.6, the encoder has access to the host data $X_i^{N_i}$ and to the uniquely assigned secret keys $K_i^H$ and $K_i^{DH}$, which are uniformly distributed over the sets $\mathcal{K}_i^H = \{1, 2, \ldots, |\mathcal{K}_i^H|\}$ and $\mathcal{K}_i^{DH} = \{1, 2, \ldots, |\mathcal{K}_i^{DH}|\}$ for robust hashing and data hiding, respectively. The secret key $K_i^H$ and the host data $X_i^{N_i}$ are used to generate a hash message $M_{3-i}$ that is encoded into the watermark $W_i^{N_i}$ based on $X_i^{N_i}$ and the secret key $K_i^{DH}$. The watermark $W_i^{N_i}$ is embedded into the host data $X_i^{N_i}$, resulting in the watermarked data $Y_i^{N_i}$. The watermarked data $Y_i^{N_i}$ is communicated through the channel $p(v_i^{N_i}|y_i^{N_i})$, which introduces some legitimate distortions.

For authenticating the document, the decoder outputs $\hat{M}_i$, an estimate of $M_i$, based on the distorted data $V_i^{N_i}$ and $K_i^{DH}$. Additionally, the hash $\hat{L}_{3-i}$ is computed from $V_i^{N_i}$ and $K_i^H$. Finally, the decision about the authenticity of $(V_1^{N_1}, V_2^{N_2})$ is made based on the comparison of $(\hat{M}_1, \hat{M}_2)$ with $(\hat{L}_1, \hat{L}_2)$. We assume that the hash message $M_i \in \mathcal{M}_i$ and the hash $L_i \in \mathcal{L}_i$ are uniformly distributed over $\mathcal{M}_i = \{1, 2, \ldots, |\mathcal{M}_i|\}$ and $\mathcal{L}_i = \{1, 2, \ldots, |\mathcal{L}_i|\}$, respectively. We also assume that $|\mathcal{M}_i| = 2^{N_i-R_{DH}^{3-i}}$ and $|\mathcal{L}_i| = 2^{N_3-R_{H}^{3-i}}$, where $R_{DH}^{3-i}$ is the data hiding rate, $R_{H}^{3-i}$ is the hashing rate, and $N_i$ is the length of all the involved vectors $X_i^{N_i}, W_i^{N_i}, Y_i^{N_i}$, and $V_i^{N_i}$.

It is straightforward to show (see Section 3.3) that if $X_1^{N_1}$ and $X_2^{N_2}$ are finite alphabet stochastic processes that satisfy the asymptotic equipartition property, then there are robust hashing and data hiding codes with specified average probability of false alarm $P_F$ and vanishing average probability of miss $P_M$ as $N_1 \to \infty$ and $N_2 \to \infty$, if the rates of the hashing and data hiding codes $R_H^{1}, R_H^{2}, R_{DH}^{1}$ and $R_{DH}^{2}$ satisfy $R_H^{1} \leq R_{DH}^{1} \leq C_2$ and $R_H^{2} \leq R_{DH}^{2} \leq C_1$. 

Figure 5.5: Proposed enrollment and production of an identification document.
Figure 5.6: Proposed authentication of an identification document.
5.5 Practical Scheme for Document Authentication

5.5.1 Information Storage with Digital Data Hiding

In order to implement the authentication system described in Section 5.4.2, we need to specify the data hiding algorithms to be used for both images (biometrics) and text data (personal data).

In this work we will use the image data hiding algorithm described in [85]. The reader is referred to the cited publication in order to obtain all the relevant implementation details.

As for the text data hiding algorithm, we will use color index modulation (CIM). As described in Section 4.4.1, this technique is fully automatable, has high information embedding rate, shows good resistance to printing and scanning, and can be applied to both electronic and printed text documents.

5.5.2 Robust Visual Hashing

According to the system architecture shown in Figure 5.3c, the protection of the identification document is based on the cross-storage of biometrics and personal data. Since biometrics represent a relatively large amount of information stored in the form of images, the storage of this information inside the personal data is a great challenge due to the low embedding rate of text data hiding techniques. Therefore, one can either compress the data and then encrypt it using proper cryptographic techniques, thus satisfying the security requirements and the rate requirements of the text data hiding submodule; or apply adequate cryptographic hashing. Both approaches share a common drawback: a high sensitivity to any change in the sense that any stream modification, even of a single pixel, causes a different result. The same is true for the personal data that should be stored inside the biometric data. Indeed, since the authentication task is performed using mobile imaging devices, which produce low quality images, the image data hiding rate is also limited and a similar problem exists. That is why there is a great need for tolerant visual hash functions that are robust to legitimate changes.

For the feature extraction from the visual input, we have to define what a “legitimate” alteration is, and which inputs can be considered as “perceptually equivalent”. This aspect concerns both the type and the level of distortion we want to allow, and is dependent on the targeted application. The allowable distortions to which the selected features should be invariant could include signal processing changes such as slight lossy compression, signal fading, noise addition, gray-scale conversion, etc., as well as some classes of geometrical distortions.

5.5.2.1 Robust Visual Hashing for Personal Data

We will consider the OCR + MAC text hashing technique discussed in Section 4.5.1. One attractive feature of this technique is that it is compatible with character feature text data hiding methods such as CIM and location index modulation (LIM). As shown in Section 4.6.2 this technique provides good robustness against legitimate modifications
and excellent detection of illegitimate modifications if OCR is applied in character-by-character mode.

5.5.2.2 Robust Visual Hashing for Biometrics

Early robust visual hashing algorithms for images have been proposed by Schneider and Chang [86], which use features like edges, color/gray-scale histograms, or discrete cosine transform (DCT) coefficients; and by Brandt and Lin [87] which are also robust to translation, rotation, and scaling. Xie and Arce [88] extract information from edges using the discrete wavelet transform (DWT). Bhattacharjee and Kutter [47] extract perceptually interesting feature points that are not embedded within the image but are stored separately. Later Hel-Or et al. [89] proposed geometric hashing based on salient points and a voting algorithm; and Fridrich [2] proposed a robust hash function based on the projections of low-frequency DCT coefficients over random patterns, which can be made invariant to translation, scaling and rotation using the Fourier-Mellin transform [90]. Another approach was considered by Monga and Evans [91], which extracts geometry preserving image features in order to generate the hash. Mihcak and Venkatesan [92] proposed an iterative robust hashing algorithm based on repeated thresholding and spatial filtering. Lefebvre et al. [93] proposed the Radon soft hash algorithm (RASH) in order to handle desynchronization and geometric distortions. Finally, Swaminathan et al. [43] developed a robust hashing algorithm based on Fourier transform features and controlled randomization.

The randomization step, generally based on a key $K_{H}$, is essential, since the generated hash function should keep the same properties as a classical cryptographic hash function besides its continuous character: hashes should be unpredictable for random inputs, and two completely different inputs should result into uncorrelated hashes. In the case of keyed hashing, two different keys should also produce totally different hashes. Fridrich [2] uses key-dependent random matrices to randomize low-pass DCT coefficients, and Venkatesan et al. [94] propose a random tiling of the DWT of the input prior to feature extraction.

The data reduction step is analogous to lossy data compression, which reduces the length of the encoded features to a compact digest code. Both the randomization and the data reduction steps should preserve the continuous property of the input features. For this purpose these two steps could be done jointly rather than separately.

The verification is then done by counting the percentage of mismatching bits with a threshold representing the amount of allowed distortion.

The design we consider closely follows the design of Fridrich [95]. Let the image $x^{N}$ be a realization of a vector random variable $X^{N}$. Let $b^{S}$ be its hash of length $S$ with key $k_{H}$, i.e. $b^{S} = \theta(x^{N}, k_{H})$. Then, $b^{S} = (b_{1}, b_{2}, \ldots, b_{S})$ is computed as follows.

i. For every $s \in \{1, 2, \ldots, S\}$, a mask $\lambda_{s}^{N} = (\lambda_{s1}, \lambda_{s2}, \ldots, \lambda_{sN})$ is generated from a uniform distribution over the open interval $]-\frac{1}{2}, \frac{1}{2}[\]$, dependent on $k_{H}$.

ii. For every $s \in \{1, 2, \ldots, S\}$, we compute:

$$\mu_{s} = \sum_{j=1}^{N} \lambda_{sj} x_{j}.$$
iii. We determine $\tau_m$ as the median of all $\mu_s$, $s \in \{1, \ldots, S\}$.

iv. Each bit $s \in \{1, 2, \ldots, S\}$ of the hash $b^S$ is determined as:

$$b_s = \begin{cases} 0 & \text{if } \mu_s < \tau_m, \\ 1 & \text{otherwise.} \end{cases}$$

5.6 Experimental Results

In this section, we present the experimental results showing the feasibility of the proposed authentication system. In order to be as close as possible to reality, we take as model the current Swiss driving licence shown in Figure 5.7. According to this card model, we assume the person’s image to be of size 320x400 pixels and the total number of text characters to be at least 200.

The different modules of the proposed system are the following:

i. OCR + MAC as robust text hashing technique,

ii. CIM as text data hiding technique,

iii. Robust image hashing as described in Section 5.5.2.2,

iv. Image data hiding as described in [85].

For the implementation of OCR + MAC text hashing we used ABBYY FineReader as OCR tool and HMAC SHA-1 truncated to 64 bits as MAC. Since the used image data hiding technique can reliably store 64 information bits, this implementation takes into account the rate requirement of the image data hiding part of the authentication system, namely that $R_H^2 \leq R_{DH}^1$.

CIM was implemented using the parameters described in Section 4.6.1. In particular, we used the quantizer set given by $Q_0(x) = 0$ and $Q_1(x) = 46$. The detection was performed by combining the mean and variance metrics of each character’s luminance as
described in [96]. In order to ensure error-free decoding we employed an outer shortened BCH code with parameters \((n, k) = (192, 68)\), which was derived from the \((255,131)\) BCH code with generator polynomial given by [97, Table 10.8]:

\[
g(D) = 215713331471510151261250277442142024165471.
\]

The used robust image data hiding algorithm is fully described in [85]. For the considered application, this algorithm reliably stores 64 information bits on an image of size \(320 \times 400\) at an image PSNR of 30 dB. All of the tested images were printed at the resolution of 300 ppi. These parameters make the watermarked image resistant to all legitimate distortions including printing on paper using halftoning or on plastic using laser engraving as well as scanning or taking a picture using the digital camera of a mobile device.

The implementation of the robust image hashing algorithm presented in Section 5.5.2.2 is straight-forward. The algorithm was tested on images of size \(320 \times 400\).

These modules were implemented separately and tested for the required conditions of document authentication. Furthermore, all modules showed very good performance when the identification document was printed and scanned using commodity printers and scanners. When using digital cameras of mobile phones consistent results were found for all the modules but the text data hiding module which suffers from the introduced blurring. As possible countermeasure we propose the use of microlenses, higher quality digital cameras, or the use LIM for text data hiding.

The experimental results for each component subsystem can be found in Sections 4.6.1 and 4.6.2, and in [85] (see also [96] for further results on the CIM method). Therefore, we only report here the obtained results for the robust image hashing module, which are new.

Hashes of different lengths have been computed from 400 images, with 10 photo’s each of 40 different subjects [98]. The Hamming distance between each possible pair of hashes of the same length has been computed. Ideally, for a given length \(S\), the Hamming distance between the hashes of unrelated images is half the number of bits in the hash, i.e. \(S/2\). Figure 5.8 shows histograms of Hamming distances between different images for hashes of different lengths. The graph clearly shows that the behavior of the hash function is approaching the ideal case with various rates that depend on the hash length. The hash can thus discriminate between images of different people.

The next step is to verify that the hash of an image can still be computed after a number of legitimate distortions. To verify this property, we have watermarked, printed and scanned 10 images [99], and then compared the output to the images without modifications. Essentially, the unmodified image is then the channel input, the channel output is the scanned image, and the channel is formed by the watermarking procedure, the printing process, and the optic equipment used for scanning. Based on the results shown in Figure 5.8 and taking into account the capabilities of the text data hiding subsystem, namely that \(R_{H}^{1} \leq R_{DH}^{2}\), we selected a hash length of 64 bits. The experiments have been repeated 100 times for different values of the key \(k_{H}\), which, due to the construction of the hash, may lead to a different Hamming distance between the hashes of any two images. A histogram of the Hamming distances between 64-bit hashes of images with an without legitimate distortions is shown in Figure 5.9, it is a delta-pulse as all these Hamming distances are zero. The histogram of Hamming distances for different images is shown for reference in the same figure. From this graph, it is clear that it is possible to define a threshold \(T\), for
Experimental Results

Figure 5.8: Hamming distances between different images.

example, $T = 2$, where images are considered equal if the Hamming distance between the hashes is smaller than $T$. These experiments show that, even with a simple approach to robust visual hashing, the security of identification documents can be enhanced by adding the proposed link between biometric and personal data, and that it is a viable approach due to its robustness.

A more detailed security analysis of the practical scheme is desirable. According to Kerckhoffs’ principle, the security of the scheme depends only on the secret key. There are four keys used in our system, one for hashing the text, one for hashing the image, one for embedding into the text and one for embedding into the image. A brute-force attack would then require the forger, in the possession of at least one valid document, to find four keys such that if both the image and text are hashed and the hashes embedded, then the end result is fully identical to the original document. Let us assume that the forger can indeed determine whether the end result is correct or not. Looking at only one of the two hash-embed sequences, observe that a correct result can only be obtained if both keys are correct, which implies that the cardinality of the combined key space is the product of the cardinalities of the individual key spaces. The messages to be embedded, which are the hashes, can be at least 64 bits, and if we assume that the key space should not be larger than the message space, this would limit the total key space to no less than 128 bits for each pair.

Since the forger’s search for the secret keys can be facilitated by various kinds of available side information, we are planning to perform in the future a thorough secu-
Authentication of Biometric Identification Documents

Figure 5.9: Hamming distances between 64-bit hashes of images before and after being watermarked, printed and scanned, which are legitimate distortions. The Hamming distances of different images are provided for reference.

Rity analysis of the system based on Shannon’s equivocation measure [100]. In order to quantify the efforts of the forger in revealing the secret keys, one can use the concept of unicity distance for the situation when multiple IDs secured with the same keys are available. This analysis should provide the exact number IDs that are necessary to reveal the secret keys without a single bit error. This analysis falls outside of the scope of this Chapter.

The practical design of an authentication system is usually subject to a trade-off between robustness and security. Robustness refers to the system’s capabilities to withstand a class of legitimate distortions. It is usually provided by introducing a certain level of redundancy via repetitive information embedding and the use of error control codes. Security can be defined as follows: only an authorized user with knowledge of the required secret key is granted the right to authenticate the data. It is guaranteed by the robust hashing subsystem, and, in terms of computational security, is measured by the length of the hash. Thus, by fixing the level of robustness, that simultaneously defines the embedding rate, one obtains the number of hash bits that can be reliably hidden. Oppositely, by fixing the number of hash bits one can deduce the number of block repetitions as well as the rate of the error control codes that equivalently define the class of distortions that the developed authentication system is guaranteed to be resilient to.
5.7 Conclusions

In this Chapter, we considered the problem of authentication of biometric identification documents via mobile devices such as mobile phones or PDAs. The proposed solution makes use of data hiding in order to cross-store the biometric data inside the personal data and vice versa. Moreover, a theoretical framework was presented which enables the analysis of future authentication systems based on this approach and guides their design. In particular, we advocate the separation approach which uses robust visual hashing techniques in order to match the information rates of biometric and personal data to the rates offered by current image and text data hiding technologies.

We also described practical schemes for robust visual hashing and data hiding that can be used as building blocks for the proposed authentication system. These schemes share a common property, namely resistance to legitimate distortions to which the identification document may be subjected. The obtained experimental results show that the proposed authentication system constitutes a viable and practical solution when the document acquisition is performed using a CCD scanner.

The main results of this Chapter have been published in [107].
6 Conclusions and Future Research Directions

6.1 Conclusions

In this thesis, we have investigated the general problem of document authentication and, in particular, of text document authentication. The fundamental requirement imposed on the systems studied in this thesis was that they should apply to both the electronic and printed forms of a document.

Firstly, due to their flexibility, low price and omnipresence, two-dimensional (2-D) bar codes were studied as auxiliary data storage means for hash-based document authentication. The main targeted problem was that of limited data storage capability of current 2-D bar codes for some document authentication applications. In the scope of this problem, multilevel 2-D bar codes were shown, for the first time, to possess the highest data storage rates in comparison to other public domain 2-D bar codes. Two main contributions were at the heart of this result. The first one was the construction of a simplified yet reasonable model for the print-scan (P-S) channel specifically adapted to the problem of data communication via multilevel 2-D symbols. The second one was the adaptation of multilevel coding / multistage decoding, a coded modulation technique originally developed for the additive white Gaussian noise channel, to the P-S channel.

Secondly, being aware that for some document authentication applications it is not possible to use a 2-D bar code, for example because of esthetical reasons, we considered data hiding as an alternative data storage technology for hash-based document authentication. In this context, we carried out an information-theoretic performance analysis of a generic document authentication system as well as justified the appropriateness of a document authentication system based on robust visual hashing and data hiding. Furthermore, by using an information equivocation argument, we also conducted a security analysis of the latter and proposed brute force attacking strategies with corresponding complexity estimates that, contrarily to existing studies, crucially rely on the information leakage of the document authentication system.

Thirdly, we studied the particular problem of text document authentication via robust text hashing and text data hiding. In this context, we proposed a novel theoretical framework for the problem of data hiding in text documents. We explained how this problem
can also be seen as an instance of the well-known Gel’fand-Pinsker problem. The main idea was to consider a text character as a data structure consisting of multiple quantifiable features such as shape, location, orientation, size, color, etc. The power of this framework was demonstrated by showing that previous text data hiding techniques, namely open space methods and character feature methods, are particular cases of a general quantization-based text data hiding scheme, which in turn is a practical implementation of the general Gel’fand-Pinsker text data hiding scheme. For illustration, we presented color index modulation as an original method for semi-fragile data hiding of electronic and printed text documents. The experimental work confirmed that this method has high perceptual invisibility, high information embedding rate, and is fully automatable. We also studied two robust text hashing algorithms, namely OCR + MAC text hashing and random tiling text hashing, that are particularly well suited for the considered authentication problem. In particular, we showed by experimentation that the former exhibits better applicability than the latter. Moreover, the experimental work also confirmed that both text hashing algorithms are robust against typical legitimate document distortions including electronic format conversion, printing, scanning, photocopying, and faxing.

Finally, we applied our theoretical and practical findings on (text) document authentication in order to design a novel authentication system for biometric identification documents, which may be used in conjunction with mobile devices such as mobile phones or PDAs equipped with a digital camera. We assumed that the biometric identification document holds biometric data (e.g. face or fingerprint) in the form of an image and personal data in the form of text, both being printed directly onto the identification document. The proposed solution made use of data hiding in order to cross-store the robust hashed biometric data inside the personal data and vice versa. The obtained experimental results showed that the proposed system constitutes a viable and practical solution.

The main results of this thesis have been published in [101, 102] regarding multilevel 2-D bar codes, in [103] regarding the information-theoretic aspects of document authentication, in particular, via robust hashing and data hiding, in [104, 105, 106] regarding practical methods for robust text hashing and text data hiding, and in [107] regarding the application of the developed technologies to the problem of authentication of biometric identification documents.

## 6.2 Future Research Directions

As possible extensions of the results obtained in the scope of this thesis, we propose the following research directions.

In the context of authentication via multilevel 2-D bar codes, one can investigate the synchronization problem for P-S channels. A good synchronization algorithm would lead to a performance improvement in terms of bit error rate (BER). One can also use irregular low density parity-check codes (instead of quasi-regular ones) and larger block lengths to improve the multistage decoder performance. It is also possible to improve the BER performance by considering skewed and non-skewed family distributions for modeling the noise distribution of the discrete P-S channel. Finally, given that a significant improvement in terms of storage rate may be obtained by eliminating the space between 2-D symbols, a more general framework including inter-symbol interference cancelation
is to be investigated.

In the scope of the information-theoretic analysis of document authentication, it would be of interest to complete the performance investigation of the considered system by studying error exponents as well as all of the related security issues. Additionally, one could also analyze the security of practical authentication systems based on the security framework developed in Chapter 3.

Concerning practical text document authentication via robust visual text hashing and text data hiding, it is desirable to provide possible countermeasures against the weaknesses of both text hashing methods as well as to perform a security analysis.

Finally, since all of the experimental results were obtained for CCD scanners and being conscious of the fact that mobile devices have become ubiquitous, it is reasonable to investigate the performance of the proposed technologies, and, eventually, modify them by taking into account the limitations of such devices, i.e. low computational power, low quality acquisition equipment, high desynchronization, etc.
Bibliography


Publications by Renato F. Villán S.


2-D two-dimensional
AEP asymptotic equipartition property
ASCII American standard code for information interchange
AWGN additive white Gaussian noise
BER bit error rate
BL bilevel
B&W black and white
CCD charged coupled device
CD compact disc
CIM color index modulation
D-A-D digital-to-analog-to-digital
dB decibel
DCT discrete cosine transform
DNA deoxyribonucleic acid
DOC Word document
dpi dots per inch
DWT discrete wavelet transform
ECC error control coding
G-P Gel’fand-Pinsker
<table>
<thead>
<tr>
<th>Acronym</th>
<th>Definition</th>
</tr>
</thead>
<tbody>
<tr>
<td>GGD</td>
<td>generalized Gaussian distribution</td>
</tr>
<tr>
<td>HIM</td>
<td>halftone index modulation</td>
</tr>
<tr>
<td>HMAC</td>
<td>hash message authentication code</td>
</tr>
<tr>
<td>HVS</td>
<td>human visual system</td>
</tr>
<tr>
<td>ID</td>
<td>identification document</td>
</tr>
<tr>
<td>ISI</td>
<td>inter-symbol interference</td>
</tr>
<tr>
<td>LDPC</td>
<td>low density parity-check</td>
</tr>
<tr>
<td>LIM</td>
<td>location index modulation</td>
</tr>
<tr>
<td>lpi</td>
<td>lines per inch</td>
</tr>
<tr>
<td>MAC</td>
<td>message authentication code</td>
</tr>
<tr>
<td>ML</td>
<td>multilevel</td>
</tr>
<tr>
<td>MLC</td>
<td>multilevel coding</td>
</tr>
<tr>
<td>MMSE</td>
<td>minimum mean-squared error</td>
</tr>
<tr>
<td>MSD</td>
<td>multistage decoding</td>
</tr>
<tr>
<td>OCR</td>
<td>optical character recognition</td>
</tr>
<tr>
<td>p.d.f.</td>
<td>probability density function</td>
</tr>
<tr>
<td>p.m.f.</td>
<td>probability mass function</td>
</tr>
<tr>
<td>PAM</td>
<td>pulse amplitude modulation</td>
</tr>
<tr>
<td>PDA</td>
<td>personal digital assistant</td>
</tr>
<tr>
<td>PDF</td>
<td>portable document format</td>
</tr>
<tr>
<td>ppi</td>
<td>pixels per inch</td>
</tr>
<tr>
<td>P-S</td>
<td>print-scan</td>
</tr>
<tr>
<td>PS</td>
<td>PostScript</td>
</tr>
<tr>
<td>PSNR</td>
<td>peak signal-to-noise ratio</td>
</tr>
<tr>
<td>pt</td>
<td>point</td>
</tr>
<tr>
<td>RASH</td>
<td>Radon soft hash algorithm</td>
</tr>
<tr>
<td>RFID</td>
<td>radio frequency identification</td>
</tr>
<tr>
<td>RH-DH</td>
<td>robust hashing and data hiding</td>
</tr>
<tr>
<td>Acronym</td>
<td>Definition</td>
</tr>
<tr>
<td>---------</td>
<td>------------</td>
</tr>
<tr>
<td>RTF</td>
<td>rich text format</td>
</tr>
<tr>
<td>SHA-1</td>
<td>secure hash algorithm 1</td>
</tr>
<tr>
<td>SM</td>
<td>storage module</td>
</tr>
<tr>
<td>SCS</td>
<td>scalar Costa scheme</td>
</tr>
<tr>
<td>SNR</td>
<td>signal-to-noise ratio</td>
</tr>
<tr>
<td>SUSAN</td>
<td>smallest unvalue segment assimilating nucleus</td>
</tr>
<tr>
<td>TEX</td>
<td>\LaTeX/\TeX document</td>
</tr>
<tr>
<td>VWP</td>
<td>Visa Waiver Program</td>
</tr>
</tbody>
</table>
List of Symbols

Chapter 1

$K_H$ secret key for the robust hashing subsystem, page 2
$K_{BC}$ secret key for the 2-D bar code subsystem, page 3
$K_{DH}$ secret key for the data hiding subsystem, page 4

Chapter 2

$r_p$ printer’s resolution measured in dpi, page 13
$r_s$ scanner’s resolution measured in ppi, page 13
$a$ length in dots of the side of a square halftone cell, page 13
$U$ number of bits that can be stored in one square inch, page 13
$M$ number of gray levels used by a multilevel 2-D bar code, page 14
$r_s^{BL}$ required scanner’s resolution for a bilevel 2-D bar code, page 14
$r_s^{ML}$ required scanner’s resolution for a multilevel 2-D bar code, page 14
$r_{im}$ 2-D bar code image resolution measured in ppi, page 15
$X$ input gray value of a 2-D symbol, page 15
$Y$ output gray value of a 2-D symbol, page 15
$J$ total number of 2-D symbols sent over the continuous P-S channel, page 16
$y_j(x)$ $j$-th received noisy 2-D symbol corresponding to the sent 2-D symbol with gray value $x$, page 19

113
\( \hat{\mu}_{Y|X}(x) \) sample mean of the received noisy symbols \( y_j(x), j = 1, \ldots, J \), page 19
\( \hat{\sigma}^2_{Y|X}(x) \) sample variance of the received noisy symbols \( y_j(x), j = 1, \ldots, J \), page 19
\( X \) signal constellation (set of gray values), page 19
\( \hat{\gamma}_{Y|X}(x) \) sample shape parameter of \( (Y|X = x) \), page 21
\( \Gamma(x) \) Gamma function, page 21
\( \rho(x) \) print-scan channel response, page 21
\( Z \) zero-mean additive noise, page 21
\( C_{\text{AWGN}} \) capacity of the AWGN channel, page 23
\( L \) number of bits required to label the signal points in \( X \), page 23
\( \psi \) bijective mapping associating signal points and binary labels, page 24
\( B^i, b^i \) \( i \)-th bit of a binary label, page 24
\( q \) binary data block, page 24
\( K \) length of the binary data block \( q \), page 24
\( q^i \) binary data subblock, page 24
\( K_i \) length of the binary data subblock \( q^i \), page 24
\( E_i \) \( i \)-th binary encoder, page 26
\( R_i \) rate of the \( i \)-th binary code, page 26
\( x \) multilevel codeword, page 26
\( x_i \) \( i \)-th symbol of \( x \), page 26
\( N \) codeword blocklength, page 26
\( R \) rate of the multilevel code, page 26
\( D_i \) \( i \)-th binary decoder, page 26
\( y \) noisy multilevel codeword, page 27
\( y_i \) \( i \)-th symbol of \( y \), page 27
\( \hat{b}^j \) estimate of \( b^j \), page 27
\( \hat{b}^j_n \) \( n \)-th symbol of \( \hat{b}^j \), page 27
\( \hat{q}^j \) estimate of \( q^j \), page 27
\( \hat{q} \) estimate of \( q \), page 27
List of Symbols

\( l^i_n \) log-likelihood ratio for the \( n \)-th bit at the \( i \)-th level, page 33

\( \sigma^2_Z \) noise variance, page 33

Chapter 3

\( K_{\text{DH}} \) secret key for the data hiding subsystem, page 39

\( \varphi \) encoder of a generic document authentication system, page 40

\( X^N \) original document, page 40

\( K \) secret key for a generic document authentication system, page 40

\( Y^N \) tamper-resistant document, page 41

\( V^N \) distorted document, page 41

\( \xi \) decoder of a generic document authentication system, page 41

\( D^E \) encoder’s allowable distortion, page 41

\( D^C \) channel’s allowable distortion, page 41

\( H_0 \) null hypothesis (the received document is authentic), page 41

\( H_1 \) alternative hypothesis (the received document is fake), page 41

\( P_F \) probability of a false alarm, page 41

\( P_M \) probability of a miss, page 41

\( P_F(k) \) probability of a false alarm given that \( K = k \), page 41

\( P_M(k) \) probability of a miss given that \( K = k \), page 41

\( T \) threshold for testing hypotheses, page 42

\( K_{\text{H}} \) secret key for the robust hashing subsystem, page 45

\( L \) robust hash, page 45

\( M \) hash message, page 45

\( W^N \) watermark, page 45

\( d_C(y^N, v^N) \) channel distortion between \( y^N \) and \( v^N \), page 45

\( d_E(x^N, y^N) \) encoding distortion between \( x^N \) and \( y^N \), page 46

\( C \) Gel’fand-Pinsker data hiding capacity, page 47

\( U \) auxiliary random variable, page 47
\[ R_{\text{DH}}' \] number of bits used to represent an individual symbol of the interference \( x^N \), page 47

\( \zeta \) deterministic mapping used to generate the watermark \( w^N \) from the code-word \( u^N \) and the original document \( x^N \), page 47

\[ d(x^N_1, x^N_2) \] distortion between \( x^N_1 \) and \( x^N_2 \), page 48

\( \theta \) robust hash function, page 48

\( D \) maximum allowable distortion, page 49

\( \Xi \) key-dependent transform, page 49

\( Q \) key-dependent quantizer, page 50

\( Q_0 \) dithered quantizer, page 50

\[ d_{N'} \] length–\( N' \) dither vector, page 50

**Chapter 4**

\( M \) message to be hidden, page 59

\( X^N \) cover text, page 59

\( W^N \) watermark, page 59

\( Y^N \) stego text, page 59

\( C_{\text{AWGN}} \) capacity of the additive white Gaussian noise channel, page 59

\( \alpha' \) compensation parameter, page 59

\( Q_m(X) \) scalar quantizer for message \( m \), page 59

\( Q_m^N(X^N) \) vector quantizer for message \( m \), page 60

\( f_{\text{HT}} \) halftone encoder, page 67

\( g_{\text{HT}} \) halftone decoder, page 67

\( \theta \) robust hash function, page 68

\( K_H \) secret key for the robust hashing subsystem, page 68

\( B^* \) binary representation of a robust hash, page 68

**Chapter 5**

\( \varphi \) encoder of a generic document authentication system, page 83
List of Symbols

\( K \) secret key for a generic document authentication system, page 83
\( X_1^{N_1} \) biometric data, page 83
\( X_2^{N_2} \) personal data, page 83
\( Y_1^{N_1} \) tamper-resistant biometric data, page 83
\( Y_2^{N_2} \) tamper-resistant personal data, page 83
\( V_1^{N_1} \) distorted biometric data, page 83
\( V_2^{N_2} \) distorted personal data, page 83
\( \xi \) decoder of a generic document authentication system, page 83
\( D_1^E \) encoder’s allowable distortion concerning the biometric data, page 83
\( D_2^E \) encoder’s allowable distortion concerning the personal data, page 83
\( D_1^C \) channel’s allowable distortion concerning the biometric data, page 83
\( D_2^C \) channel’s allowable distortion concerning the personal data, page 83
\( H_0 \) null hypothesis (the identification document is authentic), page 84
\( H_1 \) alternative hypothesis (the identification document is fake), page 84
\( P_F \) probability of a false alarm, page 84
\( P_M \) probability of a miss, page 84
\( K_{1\text{II}} \) secret key for the robust image hashing subsystem, page 84
\( K_{2\text{II}} \) secret key for the robust text hashing subsystem, page 84
\( K_{1\text{DH}} \) secret key for the image data hiding subsystem, page 84
\( K_{2\text{DH}} \) secret key for the text data hiding subsystem, page 84
\( L_1 \) robust hash of the personal data, page 84
\( L_2 \) robust hash of the biometric data, page 84
\( M_1 \) robust hash of the biometric data, page 84
\( M_2 \) robust hash of the personal data, page 84
\( W_1^{N_1} \) watermark for the biometric data, page 84
\( W_2^{N_2} \) watermark for the personal data, page 84
\( Y_1^{N_1} \) tamper resistant biometric data, page 84
\( Y_2^{N_2} \) tamper resistant personal data, page 84
<table>
<thead>
<tr>
<th>Symbol</th>
<th>Description</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>$V_1^{N_1}$</td>
<td>distorted biometric data</td>
<td>84</td>
</tr>
<tr>
<td>$V_2^{N_2}$</td>
<td>distorted personal data</td>
<td>84</td>
</tr>
<tr>
<td>$\hat{L}_1$</td>
<td>robust hash estimate of the personal data</td>
<td>84</td>
</tr>
<tr>
<td>$\hat{L}_2$</td>
<td>robust hash estimate of the biometric data</td>
<td>84</td>
</tr>
<tr>
<td>$\hat{M}_1$</td>
<td>robust hash estimate of the biometric data</td>
<td>84</td>
</tr>
<tr>
<td>$\hat{M}_2$</td>
<td>robust hash estimate of the personal data</td>
<td>84</td>
</tr>
</tbody>
</table>