

# Contents

## Part I Digital Documents

<b>1</b>	<b>Documents</b>	3
1.1	A Juridic Perspective	3
1.2	History and Trends	4
1.3	Current Landscape	5
1.4	Types of Documents	7
1.5	Document-Based Environments	10
1.6	Document Processing Needs	11
	References	12
<b>2</b>	<b>Digital Formats</b>	15
2.1	Compression Techniques	16
	RLE (Run Length Encoding)	16
	Huffman Encoding	16
	LZ77 and LZ78 (Lempel–Ziv)	18
	LZW (Lempel–Ziv–Welch)	19
	DEFLATE	21
2.2	Non-structured Formats	21
2.2.1	Plain Text	22
	ASCII	23
	ISO Latin	23
	UNICODE	24
	UTF	24
2.2.2	Images	28
	Color Spaces	28
	RGB	29
	YUV/YC <sub>b</sub> C <sub>r</sub>	29
	CMY(K)	30
	HSV/HSB and HLS	30
	Comparison among Color Spaces	30

Raster Graphics . . . . .	31
BMP (BitMaP) . . . . .	32
GIF (Graphics Interchange Format) . . . . .	34
TIFF (Tagged Image File Format) . . . . .	36
JPEG (Joint Photographic Experts Group) . . . . .	37
PNG (Portable Network Graphics) . . . . .	39
DjVu (DejaVu) . . . . .	41
Vector Graphic . . . . .	43
SVG (Scalable Vector Graphic) . . . . .	43
2.3 Layout-Based Formats . . . . .	45
PS (PostScript) . . . . .	45
PDF (Portable Document Format) . . . . .	56
2.4 Content-Oriented Formats . . . . .	59
2.4.1 Tag-Based Formats . . . . .	60
HTML (HyperText Markup Language) . . . . .	61
XML (eXtensible Markup Language) . . . . .	66
2.4.2 Office Formats . . . . .	69
ODF (OpenDocument Format) . . . . .	69
References . . . . .	70
<b>3 Legal and Security Aspects . . . . .</b>	<b>73</b>
3.1 Cryptography . . . . .	74
3.1.1 Basics . . . . .	74
3.1.2 Short History . . . . .	76
3.1.3 Digital Cryptography . . . . .	77
DES (Data Encryption Standard) . . . . .	79
IDEA (International Data Encryption Algorithm) . . . . .	80
Key Exchange Method . . . . .	81
RSA (Rivest, Shamir, Adleman) . . . . .	82
DSA (Digital Signature Algorithm) . . . . .	85
3.2 Message Fingerprint . . . . .	85
SHA (Secure Hash Algorithm) . . . . .	86
3.3 Digital Signature . . . . .	88
3.3.1 Management . . . . .	90
DSS (Digital Signature Standard) . . . . .	92
OpenPGP Standard . . . . .	93
3.3.2 Trusting and Certificates . . . . .	94
3.4 Legal Aspects . . . . .	97
3.4.1 A Law Approach . . . . .	98
3.4.2 Public Administration Initiatives . . . . .	101
Digital Signature . . . . .	101
Certified e-mail . . . . .	103
Electronic Identity Card & National Services Card . . . . .	104
Telematic Civil Proceedings . . . . .	104
References . . . . .	108

## Part II Document Analysis

<b>4</b>	<b>Image Processing</b>	113
4.1	Basics	114
	Convolution and Correlation	114
4.2	Color Representation	116
4.2.1	Color Space Conversions	117
	RGB–YUV	117
	RGB–YCbCr	117
	RGB–CMY(K)	118
	RGB–HSV	118
	RGB–HLS	119
4.2.2	Colorimetric Color Spaces	120
	XYZ	120
	L*a*b*	121
4.3	Color Depth Reduction	122
4.3.1	Desaturation	122
4.3.2	Grayscale (Luminance)	123
4.3.3	Black&White (Binarization)	123
	Otsu Thresholding	123
4.4	Content Processing	124
4.4.1	Geometrical Transformations	125
4.4.2	Edge Enhancement	126
	Derivative Filters	127
4.4.3	Connectivity	129
	Flood Filling	130
	Border Following	131
	Dilation and Erosion	132
	Opening and Closing	133
4.5	Edge Detection	134
	Canny	135
	Hough Transform	137
	Polygonal Approximation	139
	Snakes	141
	References	143
<b>5</b>	<b>Document Image Analysis</b>	145
5.1	Document Structures	145
5.1.1	Spatial Description	147
	4-Intersection Model	148
	Minimum Bounding Rectangles	150
5.1.2	Logical Structure Description	151
	DOM (Document Object Model)	151
5.2	Pre-processing for Digitized Documents	154
	Document Image Defect Models	155

5.2.1	Deskewing . . . . .	156
5.2.2	Dewarping . . . . .	157
	Segmentation-Based Dewarping . . . . .	158
5.2.3	Content Identification . . . . .	160
5.2.4	Optical Character Recognition . . . . .	161
	Tesseract . . . . .	163
	JTOCR . . . . .	165
5.3	Segmentation . . . . .	166
5.3.1	Classification of Segmentation Techniques . . . . .	167
5.3.2	Pixel-Based Segmentation . . . . .	169
	RLSA (Run Length Smoothing Algorithm) . . . . .	169
	RLSO (Run-Length Smoothing with OR) . . . . .	171
	X–Y Trees . . . . .	173
5.3.3	Block-Based Segmentation . . . . .	175
	The DOCSTRUM . . . . .	175
	The CLiDE (Chemical Literature Data Extraction) Approach . . . . .	177
	Background Analysis . . . . .	179
	RLSO on Born-Digital Documents . . . . .	183
5.4	Document Image Understanding . . . . .	184
5.4.1	Relational Approach . . . . .	186
	INTHELEX (INcremental THEory Learner from EXamples) . . . . .	188
5.4.2	Description . . . . .	190
	DCMI (Dublin Core Metadata Initiative) . . . . .	191
	References . . . . .	193

### Part III Content Processing

<b>6</b>	<b>Natural Language Processing . . . . .</b>	<b>199</b>
6.1	Resources—Lexical Taxonomies . . . . .	200
	WordNet . . . . .	201
	WordNet Domains . . . . .	202
	Senso Comune . . . . .	205
6.2	Tools . . . . .	206
6.2.1	Tokenization . . . . .	207
6.2.2	Language Recognition . . . . .	208
6.2.3	Stopword Removal . . . . .	209
6.2.4	Stemming . . . . .	210
	Suffix Stripping . . . . .	211
6.2.5	Part-of-Speech Tagging . . . . .	213
	Rule-Based Approach . . . . .	213
6.2.6	Word Sense Disambiguation . . . . .	215
	Lesk's Algorithm . . . . .	217
	Yarowsky's Algorithm . . . . .	217

6.2.7	Parsing . . . . .	218
	Link Grammar . . . . .	219
	References . . . . .	221
<b>7</b>	<b>Information Management . . . . .</b>	<b>223</b>
7.1	Information Retrieval . . . . .	223
7.1.1	Performance Evaluation . . . . .	224
7.1.2	Indexing Techniques . . . . .	226
	Vector Space Model . . . . .	226
7.1.3	Query Evaluation . . . . .	229
	Relevance Feedback . . . . .	230
7.1.4	Dimensionality Reduction . . . . .	231
	Latent Semantic Analysis and Indexing . . . . .	232
	Concept Indexing . . . . .	235
7.1.5	Image Retrieval . . . . .	237
7.2	Keyword Extraction . . . . .	239
	TF-ITP . . . . .	241
	Naive Bayes . . . . .	241
	Co-occurrence . . . . .	242
7.3	Text Categorization . . . . .	244
	A Semantic Approach Based on WordNet Domains . . . . .	246
7.4	Information Extraction . . . . .	247
	WHISK . . . . .	249
	A Multistrategy Approach . . . . .	251
7.5	The Semantic Web . . . . .	253
	References . . . . .	254
<b>Appendix A</b>	<b>A Case Study: DOMINUS . . . . .</b>	<b>257</b>
A.1	General Framework . . . . .	257
A.1.1	Actors and Workflow . . . . .	257
A.1.2	Architecture . . . . .	259
A.2	Functionality . . . . .	261
A.2.1	Input Document Normalization . . . . .	261
A.2.2	Layout Analysis . . . . .	262
	Kernel-Based Basic Blocks Grouping . . . . .	263
A.2.3	Document Image Understanding . . . . .	264
A.2.4	Categorization, Filing and Indexing . . . . .	264
A.3	Prototype Implementation . . . . .	265
A.4	Exploitation for Scientific Conference Management . . . . .	268
	GRAPE . . . . .	269
<b>Appendix B</b>	<b>Machine Learning Notions . . . . .</b>	<b>271</b>
B.1	Categorization of Techniques . . . . .	271
B.2	Noteworthy Techniques . . . . .	272
	Artificial Neural Networks . . . . .	272

	Decision Trees . . . . .	273
	$k$ -Nearest Neighbor . . . . .	273
	Inductive Logic Programming . . . . .	273
	Naive Bayes . . . . .	274
	Hidden Markov Models . . . . .	274
	Clustering . . . . .	274
B.3	Experimental Strategies . . . . .	275
	$k$ -Fold Cross-Validation . . . . .	275
	Leave-One-Out . . . . .	276
	Random Split . . . . .	276
<b>Glossary</b>		277
	Bounding box . . . . .	277
	Byte ordering . . . . .	277
	Ceiling function . . . . .	277
	Chunk . . . . .	277
	Connected component . . . . .	277
	Heaviside unit function . . . . .	277
	Heterarchy . . . . .	278
	KL-divergence . . . . .	278
	Linear regression . . . . .	278
	Run . . . . .	278
	Scanline . . . . .	278
<b>References</b>		279
<b>Index</b>		289



<http://www.springer.com/978-0-85729-197-4>

Automatic Digital Document Processing and  
Management

Problems, Algorithms and Techniques

Ferilli, S.

2011, XXVI, 297 p., Hardcover

ISBN: 978-0-85729-197-4