

Contents

Abstract	i
Resum	iii
Resumen	v
Agraïments (Acknowledgements)	vii
Preface	ix
Notation	xiii
Contents	xiv
1 Preliminaries	1
1.1 The field of a hundred names	2
1.2 Taxonomy of Keyword Spotting systems	4
1.2.1 Segmentation assumptions	4
1.2.2 Retrieved objects	7
1.2.3 Query representation	9
1.2.4 Training data	11
1.3 Information Retrieval	12
1.4 Pattern Recognition	14
1.5 Decision Theory	15
2 Probabilistic Keyword Spotting	19
2.1 Position-independent Keyword Spotting	19
2.1.1 Word-segmented image regions	23
2.2 Position-dependent Keyword Spotting	23

2.2.1	Relevance of an image column	24
2.2.2	Relevance of an image segment	28
2.2.3	Relevance of a transcript position	30
2.3	Query-by-example paradigm	32
2.3.1	Position-independent Keyword Spotting for QbE	33
2.3.2	Position-dependent Keyword Spotting for QbE	34
2.4	Segmentation-free spotting using position-dependent relevance	35
2.5	Relationship among position-dependent and indepen- dent relevance	37
2.5.1	Fréchet inequalities	38
3	The Probability Ranking Principle	43
3.1	Ranking multiple relevant images	43
3.2	Evaluation measures and optimality	45
3.2.1	Precision-at- k	46
3.2.2	Recall-at- k	46
3.2.3	Average Precision	50
3.2.4	Discounted Cumulative Gain	56
3.2.5	Normalized Discounted Cumulative Gain	58
3.3	Global and mean measures	59
4	Probabilistic Models for Handwritten Text	61
4.1	Image preprocessing	61
4.1.1	Text segmentation	62
4.1.2	Text line normalization	64
4.1.3	Feature extraction	65
4.2	Hidden Markov Models	66
4.2.1	Description	66
4.2.2	Training	69
4.2.3	Hidden Markov Models for Handwritten Text	71
4.3	Artificial Neural Networks	73
4.3.1	Description	73
4.3.2	Convolutional layers	75
4.3.3	Recurrent layers	77
4.3.4	Training	81
4.3.5	Neural Networks for Handwritten Text	85
4.4	Key differences between HMMs and NNs with CTC	87

4.5	<i>N</i> -gram Language Models	90
4.5.1	Combining the output of a neural network with a <i>n</i> -gram	92
4.6	Weighted Finite State Transducers	93
4.6.1	Description	94
4.6.2	Operations	98
4.6.3	The CTC algorithm as elementary WFST opera- tions	103
4.6.4	Lattices represented as WFST or WFSAs	104
5	Indexing for Fast Keyword Spotting	107
5.1	Indexing lexicon-based lattices	108
5.1.1	Position-independent relevance	108
5.1.2	Lexicon-based segment relevance	110
5.1.3	Lexicon-based transcript position relevance	112
5.2	The out-of-vocabulary problem	116
5.3	Indexing lexicon-free lattices	118
5.3.1	From character to word lattices	118
5.3.2	Lexicon-free segment relevance	123
5.3.3	Lexicon-free transcript position relevance	132
6	Probabilistic Interpretation of Traditional Approaches	137
6.1	HMM-filler method	137
6.2	BLSTM-CTC method	141
6.3	Distance-based methods	144
6.3.1	Distance-based density estimation	145
6.4	PHOC-based methods	153
6.4.1	PHOCNet	154
6.4.2	Probabilistic PHOCNet	156
7	Beyond Traditional Keyword Spotting	159
7.1	Multi-word spotting in handwritten documents	159
7.2	The future of Keyword Spotting with perfect transcripts	163
8	Experiments	169
8.1	Overview of the experimental setup	170
8.1.1	Databases	170
8.1.2	Statistical models for handwritten text	170

8.1.3	Evaluation protocol	171
8.2	Comparison of different relevance probabilities	172
8.3	Effect of the language model	177
8.3.1	Lexicon-based models	178
8.3.2	Lexicon-free models	181
8.3.3	Effect of the optical and prior scales	185
8.4	Effect of the training data size and augmentation	188
8.5	Correlation between Average Precision and Recognition Error	191
8.6	Results on other academic databases	192
8.6.1	George Washington	193
8.6.2	Parzival	195
8.6.3	Comparison with other published works	196
8.7	Using traditional GMM-HMM models	197
8.8	Segmentation-free evaluation	200
8.8.1	ICFHR2014 Handwritten Keyword Spotting Competition	200
8.8.2	ICDAR2015 Competition on Keyword Spotting for Handwritten Documents	205
8.9	Probabilistic interpretation of the HMM-Filler	207
8.9.1	Description	208
8.9.2	Results	209
8.10	Probabilistic interpretation of traditional distance-based systems	211
8.10.1	Description	211
8.10.2	Results	213
8.10.3	Discussion	215
8.11	Probabilistic interpretation of the PHOCNet	216
8.11.1	Description	216
8.11.2	Results	218
8.11.3	Discussion	219
8.12	Multi-word queries	221
8.12.1	Description	221
8.12.2	Results	222
8.12.3	Discussion	223
8.13	Summary	224
9	Large-scale demonstrators	227

9.1	Architecture design	227
9.1.1	Description of the servers	227
9.1.2	Description of the web client	230
9.2	Trésor des Chartes	232
9.3	Teatro del Siglo de Oro	235
9.4	The Bentham Collection	236
10	Conclusions	241
10.1	Contributions	241
10.1.1	Keyword Spotting probabilistic framework	241
10.1.2	Probabilistic models of handwritten text	242
10.1.3	Indexing algorithms based on the framework	242
10.1.4	Probabilistic interpretation of other methods	243
10.1.5	Beyond traditional and academic Keyword Spotting	243
10.2	Scientific publications	243
10.2.1	Probabilistic models of handwritten text	244
10.2.2	Keyword Spotting probabilistic framework	244
10.2.3	Keyword Spotting applications	246
10.2.4	Keyword Spotting competitions	247
10.2.5	Other Keyword Spotting works	248
10.3	Open source software	249
10.4	Future work	250
10.4.1	Stochastic definitions of relevance	250
10.4.2	Better statistical models and training	251
10.4.3	Probabilistic framework applied to other domains	251
A	Corpora	253
A.1	Bentham	253
A.1.1	ICFHR-2014 Competition on HTR	253
A.1.2	ICFHR-2014 Competition on KWS	256
A.1.3	ICDAR-2015 Competition on KWS	258
A.2	George Washington	260
A.2.1	Line-level experiments	260
A.2.2	Word-level experiments	262
A.3	IAM	263
A.4	Parzival	266
A.5	Plantas	268

CONTENTS

xix

List of Algorithms 271

List of Figures 273

List of Tables 279

Bibliography 283