Contents

Abbreviations **xix**

Contents **xxiii**

1 Introduction **1**

2 Fundamentals **5**
2.1 Data Representation 5
2.2 Artificial Neural Networks 12
2.3 Open and Closed Set Classification 25

3 Document Information Extraction And Classification Overview **29**
3.1 Document Layout: Transcending Single Pages . . . 34
3.2 Document Classification of Historical Manuscripts . 38
3.3 Information Extraction in Historical Manuscripts . . 43

4 Act Segmentation and Layout Analysis **49**
4.1 Problem Definition 50
4.2 A Whole-Book Evaluation Measure 57
4.3 Restricted Multi-page Act Segmentation 60
4.4 Multi-page Act Fine-Grained Segmentation 71
4.5 Simancas Archive Segmentation 92
4.6 Discussion 98

5 Document Classification **103**
5.1 Problem Definition 105
5.2 Feature Selection and Extraction for CBIDC 108
5.3 Open Set Classification in AHPC 109
5.4 Discussion 119

6 Information Extraction in Structured Documents **121**
6.1 Problem Definition 122
6.2 Evaluation Measures 128
6.3 Information Extraction in HisClima Tables 130
6.4 Discussion 141

7 Conclusions **143**
7.1 Scientific Outcomes 145
7.2 Projects 148
7.3 Open Source Software 148
7.4 Future Work 149

Appendices **153**

A Datasets **155**
A.1 Alcar - HOME 155
A.2 Archivo Histórico Provincial de Cádiz (AHPC) . . . 158
A.3 Hisclima 162
A.4 RCSA Dataset 165

List of Figures **169**

List of Tables **171**

Bibliography **173**