Contents

Acknowledgements / Agradecimientos i

Abstract / Resumen / Resum iii

1 Introduction 1
- 1.1 Motivation . . . 2
- 1.2 Mathematical Expression Recognition . . . 2
  - 1.2.1 Symbol Segmentation . . . 4
  - 1.2.2 Symbol Recognition . . . 5
  - 1.2.3 Structural Analysis . . . 7
- 1.3 Scientific Goals . . . 9
- 1.4 Document Structure . . . 9

2 Mathematical Expression Recognition 11
- 2.1 Introduction . . . 12
- 2.2 Statistical Framework . . . 15
- 2.3 Symbol Likelihood . . . 18
- 2.4 Structural Probability . . . 19
  - 2.4.1 2D Probabilistic Context-Free Grammars . . . 20
  - 2.4.2 Parse Tree Probability . . . 21

3 Symbol Recognition 23
- 3.1 Symbol Duration . . . 24
- 3.2 Symbol Segmentation . . . 24
  - 3.2.1 Introduction . . . 25
  - 3.2.2 Segmentation Model . . . 27
- 3.3 Handwritten Symbol Classification . . . 30
  - 3.3.1 Introduction . . . 30
  - 3.3.2 Online Features . . . 31
  - 3.3.3 Offline Features . . . 33
  - 3.3.4 Classifiers . . . 37

3.4 Printed Symbol Classification 40
3.4.1 Introduction 40
3.4.2 Feature Extraction 41
3.4.3 Classifiers 41

4 Structural Analysis 45
4.1 Introduction 46
4.2 Spatial Relationships Classification 46
4.2.1 Geometric Features: Bounding Boxes 47
4.2.2 Shape Features: Polar Histograms 49
4.3 Clustering-based Penalty 52
4.4 2D-PCFG Estimation 53
4.4.1 Viterbi Estimation 54
4.4.2 Constrained Parsing 54

5 Parsing Mathematical Expressions 57
5.1 2D-PCFG Parsing Algorithm 58
5.2 Complexity and Search Space 59
5.3 Multi-primitive Symbol Recognition 63
5.4 Training Process 64

6 The Problem of Performance Evaluation 69
6.1 Introduction 70
6.2 Performance Evaluation Metrics 72
6.2.1 Early Global Metrics 72
6.2.2 EMERS 73
6.2.3 IMEGE 74
6.2.4 Label Graphs 78
6.2.5 String Matching 81
6.3 Summary 82

7 Experimentation 85
7.1 Datasets 86
7.1.1 Printed Mathematical Expressions 86
7.1.2 Handwritten Mathematical Expressions 88
7.2 Classification of Handwritten Symbols 93
7.2.1 Classifiers Evaluation: HMM and RNN 93
7.2.2 Features Evaluation: Online and Offline 98
7.2.3 Comparison with other approaches 102
7.2.4 Summary 104

7.3 Classification of Printed Symbols . . . 107
7.3.1 Experimental setup . . . 107
7.3.2 Results and Discussion . . . 108
7.3.3 Summary . . . 110
7.4 Spatial Relationships Classification . . . 111
7.4.1 Geometric Feature Results . . . 111
7.4.2 Shape-Based Feature Results . . . 112
7.4.3 Discussion . . . 113
7.4.4 Summary . . . 115
7.5 Recognition of Handwritten Math Expressions . . . 117
7.5.1 Experimental Setup . . . 117
7.5.2 CROHME 2013 Experiments . . . 117
7.5.3 CROHME 2014 Experiments . . . 122
7.5.4 Summary . . . 125
7.6 Recognition of Printed Math Expressions . . . 127
7.6.1 Experimental Setup . . . 127
7.6.2 Results and Discussion . . . 129
7.6.3 Summary . . . 131

8 Applications 133

8.1 Open-source Software . . . 134
8.1.1 Printed Math Expression Recognition . . . 134
8.1.2 Handwritten Math Expression Recognition . . . 134
8.1.3 Features for Handwriting Recognition . . . 135
8.1.4 Image-based Evaluation . . . 135
8.2 LaTeX Transcription and Information Retrieval . . . 136
8.3 μCAPTCHA: Math-based CAPTCHA . . . 139
8.3.1 Introduction . . . 139
8.3.2 Related Work . . . 141
8.3.3 System Design . . . 144
8.3.4 Evaluation . . . 150
8.3.5 Discussion . . . 157
8.4 Layout Analysis of Structured Documents . . . 159
8.4.1 Introduction . . . 159
8.4.2 Segmentation of Structured Documents . . . 160
8.4.3 Probabilistic Graphical Models . . . 163
8.4.4 Grammatical Model . . . 165
8.4.5 Text Classification Features . . . 168
8.4.6 Evaluation . . . 170

8.4.7 Summary 176

9 Conclusions **177**

9.1 Summary 178

9.2 Scientific Contributions 181

List of Acronyms **185**

List of Figures **187**

List of Tables **193**

List of Algorithms **195**