

Understanding Mathematics: A System for the Recognition of On-Line Handwritten Mathematical Expressions

Dissertation

zur Erlangung des akademischen Grades
eines Doktors der Naturwissenschaften
im Fachbereich Mathematik und Informatik
der Freien Universität Berlin
vorgelegt von

Ernesto Tapia

Berlin
2004

Gutachter:

Prof. Dr. Raúl Rojas
Prof. Dr. Johan van Horebeek

Datum der Disputation:

10. Dezember 2004

Zusammenfassung

Die vorliegende Arbeit stellt ein System für die Online-Erkennung handgeschriebener mathematischer Formeln vor. Das System besteht aus zwei verschiedenen Komponenten, einem Klassifikator einzelner handgeschriebener Online-Symbole und einem Analysator mathematischer Strukturen.

Die Erkennung der einzelnen Symbole erfolgt mittels Support-Vektor-Maschinen. Aus unserer Experimenten ergab sich, dass unser Klassifikator gegenüber den klassischen Techniken bessere Erkennungsraten erreichte. Diese Ergebnisse wurden durch intensive Vorbearbeitung der Symbole und Suche optimaler Parameter ermöglicht. Unsere Experimente lassen den Schluss zu, dass Support-Vektor-Maschinen den Kompromiss zwischen Trainingszeit und Klassifikationsrate optimieren.

In der Arbeit wird eine neue Methode für die Online-Strukturanalyse handgeschriebener mathematischer Ausdrücke besprochen, die sich auf der Aufbau eines minimalen spannenden Baums und Symboldominanz basiert. Diese Technik ermöglicht eine natürliche Eingabe der mathematischen Formeln, d.h., die Symbole und Formeln werden ohne Beschränkungen nach der üblichen mathematischen Notation geschrieben. Unsere Methode lässt sich einfach erweitern, um andere mathematische Strukturen zu erkennen, z.B. Matrizen und andere ungewöhnliche Strukturen, wie die in der \LaTeX -Sprache definierte Struktur `\sideset{^{\ast}}_{\ast}`.

Unser Erkennungssystem wurde in der Programmiersprache Java implementiert und ist das Standard-Formelerkennungssystem des E-Kreide Systems.

Acknowledgments

I would like to thank Prof. Raúl Rojas for his knowledge and guidance. He introduced me into the area of artificial intelligence and pattern recognition by giving me a very interesting subject for my Ph.D. research.

I am deeply grateful to Prof. Dr. Johan van Horebeek. His insightful comments and suggestions helped me improve this thesis.

My research took place at the Institut für Informatik at the Freie Universität Berlin; without the use of its infrastructure, this work would not have been possible. My gratitude also goes to the E-Chalk team (Gerald Friedland, Lars Knipping, Kristian Jantz and Christian Zick) for providing technical support and assistance with all things related with E-Chalk, as well as to all members of the Artificial Intelligence Group of the institute.

I thank the Mexican National Council for Science and Technology (CONACyT) for its financial support during my research via the credit-scholarship number 154901. I would also like to thank the German Academic Exchange Service (DAAD) for their financial support during the instruction of German language and their support in solving all sorts of problems related to German bureaucracy.

I want to express my deep gratitude to my parents, brothers, and sisters for their support and love during the years I spent in Germany. I am grateful to my friends, Waldemar Barrera, Erik Cuevas, Marco A. Tagle, and Daniel Zaldívar, for their support and discussions, which made my stay in Germany much more pleasant. Finally, my deepest gratitude to Anja for her company and support, in particular during the last months of the preparation of this manuscript.

Contents

Zusammenfassung	i
Acknowledgments	ii
1 Introduction	1
1.1 Motivation	1
1.1.1 The Electronic Chalkboard	2
1.2 Characteristics of Handwritten Data	3
1.2.1 Off-Line and On-Line Data	3
1.2.2 Styles of Handwritten Data	6
1.3 Characteristics of Mathematical Notation	7
1.4 Steps for Recognition of Mathematical Notation	8
1.5 Objectives and Structure of this Thesis	10
2 Related Work	11
2.1 Introduction	11
2.2 Symbol Recognition	11
2.2.1 Segmentation	11
2.2.2 Preprocessing	13
2.2.3 Feature Extraction	13
2.2.4 Symbol Classification	14
2.3 Structural Analysis	16
2.3.1 Expression Formation	16
2.3.2 Error Correction	18
2.4 User Interfaces	19
2.4.1 The Natural Log System	19
2.4.2 Free Hand Formula Entry System	20
2.4.3 Infty Editor	21

2.4.4	MathJournal	22
3	Preprocessing Techniques for On-Line Handwriting	24
3.1	Introduction	24
3.2	Noise and Data Reduction	25
3.2.1	Smoothing	25
3.2.2	Point Clustering	25
3.2.3	Dehooking	26
3.2.4	Polygonal Approximation	28
3.2.5	Arc Length Resampling	29
3.3	Normalization	30
3.3.1	Stroke Grouping	30
3.3.2	Stroke's Direction and Order	31
3.3.3	Stroke Reduction	32
3.3.4	Size Normalization	33
3.4	Artificial Symbol Generation	33
3.5	Feature Extraction	37
3.5.1	Local Features	37
3.5.2	Global Features	38
4	Classification of On-Line Handwritten Symbols	40
4.1	Introduction	40
4.2	Classification Approaches	40
4.2.1	Bayesian Classification	42
4.2.2	Nearest Neighbors	43
4.2.3	Classification Trees	44
4.2.4	Artificial Neural Networks	46
4.2.5	Support-Vector Machines	50
4.3	Experimental Results	56
4.3.1	User-Dependent Classification	56
4.3.2	Experiments with the UNIPEN Database	58
5	Structural Analysis of Mathematical Expressions	66
5.1	Introduction	66
5.2	Structural Analysis	68
5.2.1	Symbol Regions and Symbol Attributes	68

5.2.2	Symbol Dominance	70
5.2.3	Baseline Representation of Expressions	71
5.3	MST Construction and Symbol Dominance	74
5.3.1	Construction of the Dominant Baseline	75
5.3.2	Construction of the Baseline Tree	77
5.3.3	Recognition of Matrices	79
5.4	Discussion	82
5.4.1	Extensions	82
5.4.2	Limitations of the Method	83
6	An Editor for On-Line Handwritten Mathematical Expressions	85
6.1	Introduction	85
6.2	Editor Capabilities	86
6.2.1	General Description	86
6.2.2	Editing and Correction	87
6.2.3	Recognition Actions	88
6.2.4	Page Actions	90
6.2.5	Pop-up Menus	90
6.2.6	Manipulation via Gestures	90
6.2.7	String Substitution	94
7	Conclusion	96
	Bibliography	99
	Anlagen gemäß Promotionsordnung	110