MSc. Thesis Proposal

Establishing Context Awareness for a Mathematical Expression Recognizer

Clare M. So
Student Number: 250017188
Department of Computer Science
University of Western Ontario
London, Ontario, CANADA. N6A 5B7
clare@scl.csd.uwo.ca

December 13, 2004

[1]This thesis is supervised by Dr. Stephen M. Watt.

Chapter 1

Introduction

1.1 Motivation

Recording mathematical knowledge electronically is not trivial because inputting mathematical expressions into a computer is cumbersome. Since mathematical expressions involve a large set of symbols and two-dimensional information, a normal keyboard cannot support inputting mathematics naturally. MathML [1], OpenMath [2], TeX and Computer Algebra System syntax are some existing methods to input mathematical expressions (Figure 1.1). These encodings of mathematical expressions may be verbose and involve a considerable amount of practice for one to use proficiently.

Pen-based devices, such as Tablet PCs, are ideal input devices for mathematicians to interact with mathematical knowledge because writing is a natural way for human to input mathematics. Unfortunately, existing handwriting recognizers of these devices cannot handle mathematical expressions (Figure 1.2). To facilitate the exchange of mathematical expression electronically, the ideal handwriting recognizers would analyze and store the expressions automatically upon capturing the digital ink from the devices.

There exist difficulties in recognizing handwritten mathematical expressions. First, mathematics expressions involve a large set of symbols. Non-Latin characters, such as Greek letters, are used in many fields in mathematics. Numerous operators represent different mathematical ideas. Different fonts of the same letter may have different meanings and usages. Second, a mathematical expression's structure is two-dimensional and a symbol's relative placement is crucial to the entire expressions' semantics. Associating symbols appropriately is important to have the intended meaning of the expressions to be intrepeted. Pen-based entry of mathematical expressions uses a combination of elements from writing and drawing. Lastly, handwriting is ambiguious. Without any contextual information, the identity of symbols and their relative symbol placement may not be determined (Figure 1.3).

In natural language handwriting recognition, a dictionary in and contextual information are needed to eliminate inadmissible results because handwriting is ambiguious. In other words, the recognizer needs certain hints to determine the result of recognition. For ex-

```
<math xmlns=
  "http://www.w3.org/1998/Math/MathML">
  <apply>
    <int/>
    <bvar>
      <ci>x</ci>
    </bvar>
    <ci>x</ci>
  </apply>
</math>
```

(a) Content MathML

```
<math xmlns=
  "http://www.w3.org/1998/Math/MathML">
  <mrow>
    <mo>&Integral;</mo>
    <mi>x</mi>
    <mo>&InvisibleTimes;</mo>
    <mrow>
      <mo>&DifferentialD;</mo>
      <mi>x</mi>
    </mrow>
  </mrow>
</math>
```

(b) Presentation MathML

```
<OMOBJ xmlns="http://www.openmath.org/OpenMath"
  version="2.0" cdbase="http://www.openmath.org/cd">
  <OMA>
    <OMS cd="calculus1" name="int"/>
    <OMBIND>
      <OMBVAR>
        <OMV name="x"/>
      <OMBVAR>
      <OMV name="x"/>
    </OMBIND>
  </OMA>
</OMOBJ>
```

(c) OpenMath

```
$\int \! x{dx}$
```

(d) TeX

```
int(x,x);
```

(e) Maple (Computer Algebra System)

Figure 1.1: $\int xdx$ in five different formats

Figure 1.2: Handwriting recognizer in Tablet PC cannot recognize mathematics

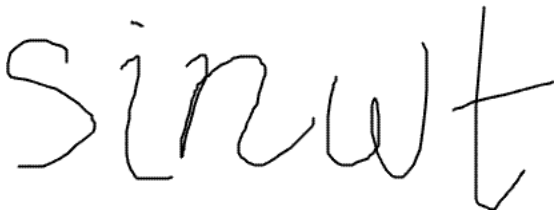

Figure 1.3: $sin\omega t$: Since ω is used with a trigonometry function, it should be interpreted as the lower case Greek letter omega instead of the Latin letter w.

ample, if the recognizer expects the users to enter an email address, it would try to locate some writing that is close to "@" and bias the result of this writing to that character. In the study of mathematical handwriting recognition, no dictionary exists to capture contextual information of different areas of mathematics.

1.2 Scope of the Project

There are several sub-problems in Mathematical Handwriting Recognition defined so far: character recognition, symbol association, digital ink architecture integration and contextual information collection. Character recognition concerns with appropriate and effective methodology to determine probable results among the large set of character in mathematics. Symbol association concerns with meaningful groupings of different characters and symbols in a mathematical expression. Ink architecture integration concerns with having the handwriting stored electronically in the same form across different handwriting devices [9] [14]. Contextual information collection conerns with examining characteristics of mathematical expressions used by actual mathematicians.

This project will be focusing on collecting, organizing and analyzing contextual information of mathematics to aid expression recognition. Different areas of advanced mathematics are examined. This part serves as the dictionary for mathematics and it does not work in isolation among the other sub-problems: By knowing patterns of symbol usages, inadmissible recognition results in character recognition and symbol association can be eliminated from the list of probable results. In other words, contextual information can restrict the results of character recognition and symbol association.

1.3 Previous and Related Works

There are some attempts in putting printed or handwritten mathematical expressions into electronic form. None of these attempts emphasizes on handling a large set of mathematical expression. In other words, each of these attempts restricts the types of mathematical expressions to be handled for its own purpose. Berman and Fateman [5] focus on encoding integral tables. Lavirotte [11] presents how graph grammar can help associating characters and symbols in formulae. Wan [13] develops an experimental mathematical handwriting recognizer for Pocket PC that is for a small set of mathematical expressions.

Searching, retriving and organizing contextual information of mathematics is a new field of study. Cairns [7] describes how Latent Semantic Indexing can help to index and retrieve information from a library of formal mathematics. This approach emphasizes on defininig expressions' semantics based on repeated occurences of sub-expressions and does not need any external ontologies to be defined.

Chapter 2

Methodology

2.1 Data Collection

TeX source and PDF files of 19389 articles on mathematics from the ArXiv e-Print archive [3] dated from January 2000 to July 2004 have been collected. ArXiv is a scientific articles archive service owned, operated and funded by Cornell University. University academic standard is maintained throughout the contents of archive. In this way, the sample mathematical expressions used by actual mathematicians from various fields are collected.

Most of the articles collected are assigned, by the authors, one or more category in Mathematical Subject Classification (2000) [4] established by American Mathematical Society. Three levels of classification are available. For this project, only the top level classification is used. Each of the 63 categories in the top level classification corresponds to a discipline of mathematics (see Appendix A). For example, category 05 corresponds to combinatorics.

2.2 Getting the Logical Structure of Expressions

For this experiment, it is important to extract the logical structure of mathematical expressions before analyzing them. TeX has been the standard for typesetting mathematical expressions, but it shows none or little logical structure. Presentation MathML exhibits richer tree-like logical structure than TeX (see Figure 2.1), but this standard is relatively new. To collect the logical structure of mathematical expressions, it is necessary to translate the expressions from TeX to Presentation MathML. The TeX to MathML converter developed by ORCCA is used.

```
${\int x}^{{2}{2} dx}$
```

```
<math xmlns=
  "http://www.w3.org/1998/Math/MathML">
  <mrow>
    <mo>&Integral;</mo>
    <msup>
      <mi>x</mi>
      <mn>22</mn>
    </msup>
    <mo>&InvisibleTimes;</mo>
    <mrow>
      <mo>&DifferentialD;</mo>
      <mi>x</mi>
    </mrow>
  </mrow>
</math>
```

Figure 2.1: Presentation MathML exhibits logical structure of expressions. The digits in number 22 should be grouped together and identify themselves as a single number. In Mathml, <mi>s indicate operators and <mn>s indicate numbers. There is no equivalent mechanism in TeX. <mrow>s indicate groupings of symbols.

2.3 Identifying Characteristics of Categories

After converting mathematical expressions from TeX to Presentation MathML, statistics on frequency of particular symbols and sub-expressions in each category are gathered. These statistics can possibly help distinguishing the characteristics of mathematics from one field to another one.

2.3.1 Letters, Greek Alphabets and Operators

Certain letters, Greek alphabets and operators are used to represent mathematical concept. Choice of these symbols is based on convention. For example, m usually represents a variable for an integer and e usually represents the Euler constant. Different fonts of the same letter is relevent on representing different mathematical concepts. π is the special constant for $3.1415...$ and $\prod$ is the projection operator. i can be the special constant $\sqrt{-1}$ for complex number or i, together with j and k, represents the unit vector.

In MathML, single-letter identifiers are enclosed by <mi>s and the operators are enclosed by <mo>s. MathML supports numerous fonts that are relevent: normal, bold, italic, bold-italic, double-struck, bold-fraktur, script, bold-script, fraktur, sans-serif, bold-sans-serif, sans-serif-italic, sans-serif-bold-italic and monospace.

After obtaining these statistics, a measure (as in Measure Theory) can be built based on the histogram on frequency of symbols in a certain category. A measure serves as a

fingerprint describing the characteristics of symbol usage in a sub-field of mathematics.

2.3.2 Common Expressions

There may be a set of commonly used expressions in a particular category. For each category, we keep track of the most common of certain size. In this experiment, we limit the size of the expressions in MathML to a certain threshold. Note that MathML markup of mathematical expression exhibits a tree-like structure (Figure 2.1). Letters, Greek letters and operators in the previous experiement are also considered in this experiement.

2.3.3 Patterns of Expressions

To extend the notion of keeping track of common expressions in a particular category, we would like to keep track of the most common patterns of expressions. In this way, common usage of certain symbols can be generalized. For example, $\sqrt{a^2 + b^2}$ and $\sqrt{x^2 + y^2}$ are considered to be two distinct expressions in the previous experiment. In this experiement, these two expressions are considered to have the same pattern of $\sqrt{A^2 + B^2}$. The formulation of a precise notion of pattern will be a part of the thesis.

Chapter 3

Conclusion

Interacting with mathematical knowledge has been challenging because inputting and editing mathematical expressions is cumbersome. Handwriting is a natural way to interact with two-dimensional information such as mathematics. Although more pen-based devices become available, these devices cannot be used to enter mathematics.

There were numerous attempts in building a handwriting recognizer for mathematics. These attempts had limited success because the set of mathematics considered was restricted. To build an improved handwriting recognizer for mathematics, it is necessary for the recognizer to know mathematics. In this way, the accuracy of recognition for large set of mathematics can be improved since inadmissible results are eliminated when the handwriting is analyzed.

The goal of this study is to extend the set of mathematics that the handwriting recognizer to be handled. We are trying to build a database describing the characteristics of mathematical expressions from various fields in advanced mathematics. In this study, articles written by actual mathematicians are collected. These articles are sorted to the different sub-fields of mathematics. Expressions from these articles are analyzed. The characteristics of a certain field of mathematics are defined upon the frequency of letters or sub-expressions appearing in the articles.

Appendix A

Mathematical Subject Classification (2000)

Category	Description
00	General
03	Mathematical logic and foundations
05	Combinatorics
06	Order, lattices, ordered algebraic structures
08	General algebraic systems
11	Number theory
12	Field theory and polynomials
13	Commutative rings and algebras
14	Algebraic geometry
15	Linear and multilinear algebra; matrix theory
16	Associative rings and algebras
17	Nonassociative rings and algebras
18	Category theory; homological algebra
19	K-theory
20	Group theory and generalizations
22	Topological groups, Lie groups
26	Real functions
28	Measure and integration
30	Functions of a complex variable
31	Potential theory
32	Several complex variables and analytic spaces
33	Special functions
34	Ordinary differential equations
35	Partial differential equations
37	Dynamical systems and ergodic theory

39	Difference and functional equations
40	Sequences, series, summability
41	Approximations and expansions
42	Fourier analysis
43	Abstract harmonic analysis
44	Integral transforms, operational calculus
45	Integral equations
46	Functional analysis
47	Operator theory
49	Calculus of variations and optimal control; optimization
51	Geometry
52	Convex and discrete geometry
53	Differential geometry
54	General topology
55	Algebraic topology
57	Manifolds and cell complexes
58	Global analysis, analysis on manifolds
60	Probability theory and stochastic processes
62	Statistics
65	Numerical analysis
68	Computer science
70	Mechanics of particles and systems
74	Mechanics of deformable solids
76	Fluid mechanics
78	Optics, electromagnetic theory
80	Classical thermodynamics, heat transfer
81	Quantum theory
82	Statistical mechanics, structure of matter
83	Relativity and gravitational theory
85	Astronomy and astrophysics
86	Geophysics
90	Operations research, mathematical programming
91	Game theory, economics, social and behavioral sciences
92	Biology and other natural sciences
93	Systems theory; control
94	Information and communication, circuits
97	Mathematics education

Bibliography

[1] David Carlisle, Patrick Ion, Robert Miner, Nico Poppelier, Editors. *Mathematical Markup Language (MathML) Version 2.0 (Second Edition)*. W3C Recommendation. `http://www.w3.org/TR/2003/REC-MathML2-20031021/`. October 21, 2003.

[2] *OpenMath Society*. `http://www.openmath.org`

[3] *ArXiv e-Print Archive*. `http://xxx.lanl.gov`
(Mirror sites: `http://www.arxiv.org`, `http://aps.arxiv.org`)

[4] *Mathematical Subject Classification (2000)*. American Mathematical Society. `http://www.ams.org/msc`

[5] Benjamin P. Berman, Richard J. Fateman. *Optical Character Recognition for Typeset Mathematics*. International Symposium on Symbolic and Algebraic Computation 1994 (ISSAC'94). July 20-22, 1994.

[6] Dorothea Blostein and Ann Grbavec. *Recognition of Mathematical Notaion*. Handbook on Optical Character Recognition and Document Image Analysis. P.S.P. Wang and H. Bunke, Editors. World Scientific Publishing Company, 1996.

[7] Paul Cairns. *Informalising Formal Mathematics: Searching the Mizar Library with Latent Semantics*. Thrid International Conferences, Mathematical Knowledge Management 2004. Biálowieza, Poland. September 2004.

[8] Yvonne Choquet-Bruhat, Cécile de Witt-Morette, Margaret Dillard-Bleick. *Analysis, Manifolds and Physics*. North-Holland Publishing Company. 1977.

[9] Kevin Durdle, *Supporting Mathematical Handwriting Recognition through an Extended Digital Ink Framework*. MSc. Thesis (Submitted). University of Western Ontario. December 2004.

[10] Paul R. Halmos. *Measure Theory*. Springer-Verlag New York Inc. 1974.

[11] Stéphane Lavirotte. *Reconnaissance structurelle de formules mathématiques typographieées et manuscrites*. PhD. Thesis. Université de Nice - Sophia Antipolis.

École Doctorale des Sciences et Technologies de l'Information et de la Communication. Institut National de Recherche en Informatique et Automatique (INRIA). June 14, 2000.

[12] J.C. Taylor. *An Introduction to Measure and Probability*. Springer-Verlag New York Inc. 1997.

[13] Bo Wan. *An Interactive Mathematical Handwriting Recognizer for the Pocket PC*. MSc. Thesis. University of Western Ontario. December 2001.

[14] Xiaojie Wu. *Achieving Interoperability of Pen Computing with Heterogeneous Devices and Digital Ink Formats*. MSc. Thesis (Submitted). Unversity of Western Ontario. December 2004.