Hierarchical Contextual Parsing Based Mathematical Formulae Recognition

Rahul Dashora

Advisor: Dr. Richard Zanibbi

Computer Science @ RIT
B.Thomas Golisano College of Computing & Information Sciences

Goal

Adapt method to recognize formulae from PDF documents, and improve parsing of extracted expressions using symbol labels where the locations of symbols are known.

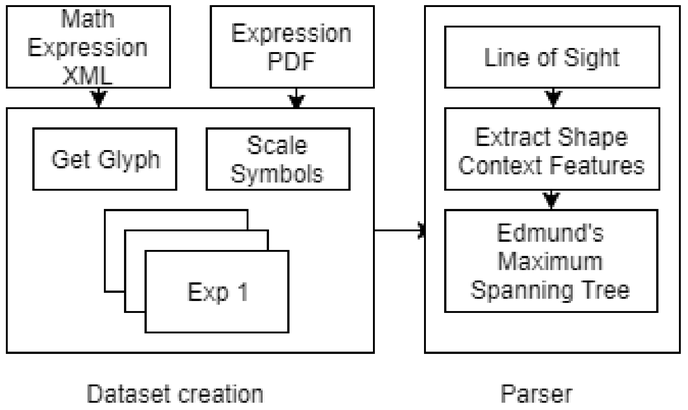
Hierarchical Contextual Parsing

Segmentation → Parsing → Classification

- **HCP – Starts with segmenting each symbol using visual , spatial and contextual features. Once segmented new set of such features are extracted, one set is used for symbol classification and another is used for parsing.**

- **Symbol Classification -** It is the task of identifying and labelling each symbol in the expression.
- **Symbol Segmentation –** This stage finds symbols within an expression. It checks various components to see if they are symbol individually or they combine in some form to represent the symbol.
- **Expression Parsing -** Take the components of the expression and output the structure of the expression in form of Symbol Layout Tree. While parsing system tries to recognize the relation between various pairs of symbols. These relations are based on the spatial orientation of symbols within the expression.
- In the previous system symbol labels had no role in parsing. But now we use that information for Syntax Context features for training the classifier for Parser.

Procedure



Syntax Context Features

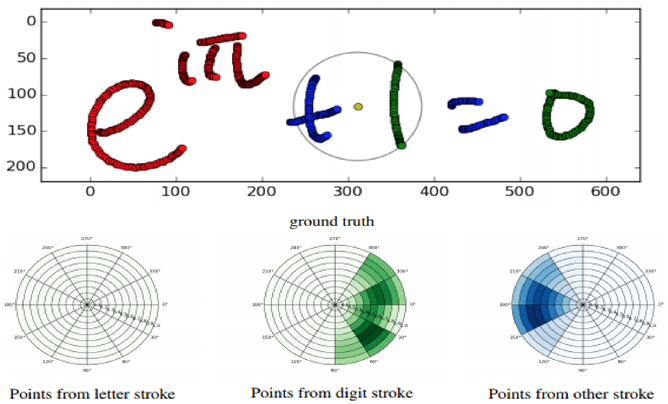
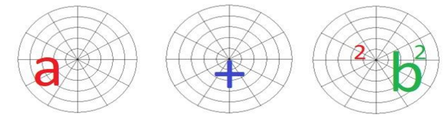


Fig. Context Syntax Features, Adapted from "Features and Algorithms for Visual Parsing of Handwritten Mathematical Expressions" by Hu Lei, 2016, Accessed from http://scholarworks.rit.edu/theses/9042

These features can improve classification because they can help in recognizing similarity in behavior between a pair of symbols of certain categories like number and operators

Visual Features

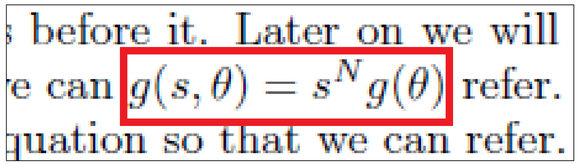
$a^2 + b^2$



Polar histograms to extract features of pairs of symbols after pruning of edges using Line of Sight graph.

Results

Expression PDF



Expression Image

$$g(s,\theta) = s^N g(\theta)$$

Symbol Lavout Tree

Conclusion

- We have successfully implemented PDF math expression parsing for the formulas extracted by MathScraper tool.

References
- Hu Lei, "Features and Algorithms for Visual Parsing of Handwritten Mathematical Expressions" (2016).
- Michael Condon "Applying Hierarchical Contextual Parsing with Visual Density and Geometric Features to Typeset Formula Recognition" (2017)