Intelligent Combination of Structural Analysis Algorithms:
Application to Mathematical Expression Recognition

APPROVED BY

SUPERVISING COMMITTEE:

Dr. Richard Zanibbi, Supervisor

ABC, Reader

XYZ, Observer

Intelligent Combination of Structural Analysis Algorithms: Application to Mathematical Expression Recognition

by

Amit Pillay

THESIS

Presented to the Faculty of the Department of Computer Science

Golisano College of Computer and Information Sciences

Rochester Institute of Technology

in Partial Fulfillment

of the Requirements

for the Degree of

Master of Science

Rochester Institute of Technology

August 2017

Acknowledgments

I wish to thank the multitudes of people who helped me. Time would fail me to tell of ...

Abstract

Intelligent Combination of Structural Analysis Algorithms: Application to Mathematical Expression Recognition

Amit Pillay, M.S.
Rochester Institute of Technology, 2017

Supervisor: Dr. Richard Zanibbi

This document has the form of a "fake" doctoral dissertation in order to provide an example of such, but it is actually a copy of Miguel Lerma's documentation for the Mathematics Department Computer Seminar of 25 March 1998 updated in July 2001 and following by Craig McCluskey to meet the March 2001 requirements of the Graduate School.

This document and its source file show to write a Doctoral Dissertation using LaTeX and the utdiss2 package.

Table of Contents

List of Tables

List of Figures

Chapter 1

Introduction

Chapter 2

Background

Mathematical Expressions MEs form an essential part of scientific and technical documents. Mathematical Expressions can be typeset or handwritten which uses two dimensional arrangements of symbols to transmit information. Recognizing both form of mathematical expressions are challenging. A variation to handwritten ME is cursive handwriting. Unconstrained cursive property of such handwritten expressions poses a major challenge to its recognition.

Generally speaking understanding and recognizing mathematical expression, whether typeset or handwritten, involves three activities: Expression localization, symbol recognition and symbol-arrangement analysis. ME localization involves finding and extracting mathematical expression from the document. Symbol recognition converts the extracted expression image into a set of symbols and symbol arrangement analyzes the spatial arrangement of set of symbols to recover the information content of the given mathematical notations.

Now based on the recognition process, symbol recognition activity can further subdivided as 1) preprocessing - noise reduction, deskewing, slant cor-

rection etc, 2) segmentation to isolate symbols 3) and finally, recognition. Similarly depending upon the symbol-arrangement algorithm, symbol arrangement analysis can be further subdivided into a) identification of spatial relationships among symbols b) identification of logical relationships among symbols 3) construction of meaning. These processes can be executed in series or in parallel with latter processes providing contextual feedback for the earlier processes. The order of these recognition activities can vary somewhat, for example, partial identification of spatial and logical relationships can be performed prior to symbol recognition.

2.1 Preprocessing

Preprocessing is required to eliminate irregularities and noise from the image, especially in handwritten character recognition. Certain preprocessing method requirements may depend upon the techniques used for recognition. [9] uses chain code method for handwritten image representation. Preprocessing involves slant angle correction in which global slant angle from different vertical lines is estimated and tangent of the estimated global slant angle is used to correct for slant. Smoothing of image involves elimination of small blobs (noise) on the contour. A sliding 3-component one dimensional window is applied overall components during which components are removed or added based on the orientation of components. Average stroke width is estimated by dividing chain code contours horizontally and by tracing left to right various distances between outer and inner contour. [10] performs size normalization

to reduce variation in character size. To avoid significant deformation due to directly scaling of all images to identical size, a holistic approach is used for scaling in which if width/height ratio is less than 0.8 then scale is identical horizontally and vertically otherwise the scale factor is set to 0.8 to prevent large variation in image width.

2.2 Character segmentation

Character segmentation, next step in ME recognition, has long been a critical area of OCR process. Depending upon the requirement, character segmentation techniques is divided into four major headings [15]. Classical approach of segmentation also called dissection technique consists of partitioning the input image into sub-images based on their inherent features, which are then classified. Another approach to segmentation is a group of techniques that avoids dissection and segments to image either explicitly by classification of pre-specified windows, or implicitly by classification of subsets of spatial features collected from the image as a whole. Another approach is a hybrid approach employing dissection but using classification to select from admissible segmentation possibility. Finally holistic approach avoids segmentation process itself and performs recognition entire character strings.

Various techniques have been used for segmentation that involves dissection. White spaces between the characters are used to detect segmentation points. Pitch which is the number of characters per unit of horizontal distance provides a basis for estimating segmentation points. The segmentation

points obtained for a given line should be approximately equally spaced at the distance that corresponds to pitch [15].

Inter-character boundaries can be obtained if most segmentation takes place by finding columns of white. Now all segmentation points that do not lie near these boundaries can be rejected as caused due to broken characters. Similarly we can estimate missed points due to merged characters. Hoffman and McCullough gave a framework for segmentation that involves three steps i.e. 1) Detection of the start of the character, 2) A decision to begin testing for the end of a character called sectioning, 3) Detection of end-of-character. Sectioning is done by weighted analysis of horizontal black runs completed versus run still incomplete. Once sectioning determines the regions of segmentation, rules were invoked to segment based on either an increase in bit density or the use of special features designed to detect end-of-character.

In [1], segmentation in cursive handwritten characters is performed in the binary word image by using the contour of the writing. Determination of segmentation regions is done in three steps. In first step a straight line is drawn in the slant angle direction from each local maximum until the top of the word image. While going upward in the slant direction, if any contour pixel is hit, this contour is followed until the slope of the contour changes to the opposite direction. An abrupt change in the slope of the contour indicates an end point. A line is drawn from the maximum to the end point and path continues to go upward in slant direction until the top of the word image. In step 2, a path in the slant direction from each maximum to the lower baseline, is drawn. Step 3

follows the same process as in step 1 in order to determine the path from lower baseline to the bottom of the word image. Combining all the three steps gives the segmentation regions. In [9] segmentation involves detecting ligatures as segmentation points in cursive scripts. Alternatively, concavity features in the upper contour and convexities in the lower contour are used in conjunction with ligatures to reduce the number of potentials segmentation points.

Another dissection technique that applies to non-cursive characters is bounding box technique [15]. In this analysis, the adjacency relationships between characters are tested to perform merging or their size or aspect ratios are calculated to trigger splitting mechanisms. Another involves splitting of connected components. Connected components are merged or split according to rules based on height and width of the bounding boxes. Intersection of two characters can give rise to special image features and different dissection methods have been developed to detect these features and to use them in splitting a character string images into sub-images.

[6] focuses on segmentation of single and multiple touching character segmentation. [6] proposes a new technique that links the feature points on the foreground and background alternately to get the possible segmentation path. Mixture Gaussian probability function is determined and used to rank all the possible segmentation paths. Segmentation paths construction is performed separately for single touching characters and for multiple touching characters. All the paths from to two analysis are collectively processed to remove useless strokes and then mixture Gaussian probability function is applied to decide

which on is the best segmentation path.

Another kind of approach to character segmentation is recognition based approach. In these segmentation processes letter segmentation is a by-product of letter recognition. The basic principle is use a mobile window of variable width to provide sequences of tentative segmentation which are confirmed (or not) by character recognition. A technique called Shortest Path Segmentation selects the optimal combination of cuts from the predefined set of candidate cuts that construct all possible legal segments through combination. A graph whose nodes represent acceptable segments is the created. The paths of these graphs represent all legal segmentations of the word. Each node of the graph is then assigned a distance obtained by the neural net recognizer. The shortest path though the graph thus corresponds to the best recognition and segmentation of the word. An alternative method attempts to match subgraphs of features with predefined character prototypes. Different alternative are represented by a directed network whose nodes correspond to the matched subgraphs. Word recognition is done by searching for the path that gives the best interpretation of the word features.

2.3 Symbol-Arrangement Analysis

One approach to symbol-arrangement analysis is syntactic approach. Syntactic approach makes use of two dimensional grammar rules to define the correct grouping of math symbols. Co-ordinate grammar for recognition is presented by Anderson. The grammar specifies syntactic rules that subdivide the

set of symbols into several subsets, each with its own syntactic subgoal. The final interpretation result is given by the m attribute of the grammar start's symbol where m represents ASCII encoding of the meaning of symbol-set. Although coordinate grammar provides a clear and well structured recognition approach, its slow parsing speed and difficulty to handle errors are its major drawbacks. In [8], a syntactic approach is adopted in which a system consisting of hierarchy of parsers for the interpretation of 2-D mathematical formulas is described. The ME interpreter consists of two syntactic parser top-down and bottom-up. It starts with a priority operator in the expression to be analyzed and tries to divide it into sub-expressions or operands which are then analyzed in the same way and so on. The bottom-up parser chooses from the starting character and from the neighboring sub-expressions the corresponding rule in the grammar. This rule gives instructions to the top-down parser to delimit the zones of neighboring operands and operators.

Garain and Chaudhari in [8], proposes a two pass approach to determine arrangement of symbols. The first pass is a scanning or lexicon analysis that performs micro-level examination of the symbols to determine the symbol groups and to determine their categories or descriptors. The second pass is parsing or syntax analysis that processes the descriptors synthesized in the first pass to determine the syntactical structure of the expression. A set of predefined rules guides the activities in both the passes.

Another symbol-arrangement analysis approach is projection profile cutting. It involves recursive projection-profile cutting. Cutting by the ver-

tical projection profile is attempted first, followed by horizontal cuts for each resulting regions. The process repeats until no further cutting is possible. The resulting spatial relationships are represented by a tree structure. Although the method looks simple and efficient technique, it is still under study and also involves additional processing for symbols like square root, subscripts and superscripts as these can be handled by projection profile cut.

Another approach discussed is the Graph Rewriting. Graph rewriting involves information represented as an attributed graph and the graph get updated through the application of graph-rewriting rules. An initial graph contains one node to represent each symbol, with nodes attributes recording the spatial coordinates of the symbol. Graph rewriting rules are applied to add edges representing meaningful spatial relationships. Rules are further applied to prune or modify these edges identifying logical relationships from the spatial relationships. In [7], Ann Grbavec and Dorothea Blostein proposed a novel-graph rewriting techniques that addresses the recursive structure of mathematical notations, the critical dependence of the meaning upon operator precedence and the presence of ambiguities that depends upon global context. The recognition system proposed called EXPRESSO, is based on Build-Constrain-Rank-Incorporate model where the Build phase constructs edges to represent potentially meaningful spatial relation- ships between symbols. The Constrain phase applies information about the notational conventions of mathematics to remove contra- dictions and resolve ambiguities. The Rank phase uses information about the operator precedence to group symbols into

sub-expressions and the Incorporate phase interprets sub-expressions.

Twaakyondo and Okamoto [18] discuss two basic strategies to decide the layout of structure of the given expression. One strategy is to check the local structures of the sub-expressions using a bottom-up method (specific structure processing). It is used to analyze nested structures like subscripts, superscripts and root expressions. The other strategy is to check the global structure of the whole expression by a top-down method (fundamental structure processing). It is used to analyze the horizontal and vertical relations between sub-expressions. The structure of the expression is represented as a tree structure.

Chou in [11] proposed a two-dimensional stochastic context-free grammar for recognition of printed mathematical expressions. The recognized symbols are parsed with the grammar in which each production rule has an associated probability. The main task of the process is to find the most probable parse tree for the input expression. The overall probability of a parse tree is computed by multiplying together the probabilities for all the production rules used in a successful parse.

2.4 Conclusion

As we saw through the survey, there have been tremendous advances in the field of character recognition from so many years of research. Some experiment tried to focus on one activity of recognition process while other tried to build a complete system for character recognition. Some researchers

assumed complete well recognized symbols are given and they focus on the symbol-arrangement (structural) analysis of the recognized symbols. This survey concentrated mainly on the two activities of character recognition i.e. segmentation of symbols and symbol-arrangement of recognized symbols.

Segmentation processes discussed have some limitations such as some are restricted to be applied to cursive handwriting while other focuses on non-cursive handwriting. Some researchers focus on certain subset of mathematical symbols because of large mathematical symbol set. Some concentrate on single touching characters some on multiple touching characters. Certain approaches of segmentation like holistic approach that recognizes entire word as a unit have drawback of being restricted to predefined lexicons. Hence more efficient and robust segmentation process is required as further analysis of ME recognition depends on segmentation and recognition of symbols.

Symbol-arrangement analysis discussed shows wide variations in approaches. Some approach exploits the operator precedence property of mathematical expression while some performs different level of analysis (lexicon and syntax) to first group symbols into different categories and then perform structural analysis using predefined rules. Some using graph rewriting technique in which mathematical symbols are linked to each other through graph rewriting rules. Some use stochastic grammar rules to represent to the relationship between symbols while some intelligently looks for local structures of the expression to determine the features like nested, above or below followed by global analysis to check for the correctness of the expression as a whole and

rectify wrong arrangements of symbols. Symbol arrangement analysis may be not so crucial for problems that involve only standard English character but problems like recognition of mathematical expressions where the actual position and location of symbols is important and there are many implicit meaning to symbols which depends on their arrangement, it is absolutely important to perform symbol arrangement analysis for better recognition result.

Chapter 3

Methodology

Chapter 4

Results and Discussion

Chapter 5

Conclusion and Future Work

Bibliography

[1] N. Arica and F.T. Yarman-Vural. Optical Character Recognition for Cursive Handwriting. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, pages 801–813, 2002.

[2] D. Blostein and A. Grbavec. Recognition of Mathematical Notation. *Handbook of Character Recognition and Document Image Analysis*, pages 557–582, 2001.

[3] RG Casey, E. Lecolinet, I.B.M.A.R. Center, and CA San Jose. A survey of methods and strategies in character segmentation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 18(7):690–706, 1996.

[4] K.F. Chan and D.Y. Yeung. An efficient syntactic approach to structural analysis of on-line handwritten mathematical expressions. *Pattern Recognition*, 33(3):375–384, 2000.

[5] K.F. Chan and D.Y. Yeung. Mathematical expression recognition: a survey. *International Journal on Document Analysis and Recognition*, 3(1):3–15, 2000.

[6] Y.K. Chen and J.F. Wang. Segmentation of Single-or Multiple-Touching Handwritten Numeral String Using Background and Foreground Analysis.

IEEE Transactions on Pattern Analysis and Machine Intelligence, pages 1304–1317, 2000.

[7] Ø. Due Trier, A.K. Jain, and T. Taxt. Feature extraction methods for character recognition-A survey. *Pattern Recognition*, 29(4):641–662, 1996.

[8] U. Garain and BB Chaudhuri. Recognition of Online Handwritten Mathematical Expressions. *IEEE Transactions on Systems, Man, and Cybernetics, Part B: Cybernetics*, 34(6):2366–2376, 2004.

[9] K. Gyeonghwan and V Govindraju. A Lexicon Driven Approach to Handwritten Word Recognition for Real-Time Applications. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, pages 366–379, 1997.

[10] Zhi-Qiang Liu Jinhai Cai. Integration of Structural and Statistical Information for Unconstrained Handwritten Numeral Recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 21(3):263–270, 1999.

[11] C.L. Liu, M. Koga, and H. Fujisawa. Lexicon-Driven Segmentation and Recognition of Handwritten Character Strings for Japanese Address Reading. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, pages 1425–1437, 2002.

[12] Y. Lu. Machine printed character segmentation-; An overview. *Pattern Recognition*, 28(1):67–80, 1995.

[13] S. Madhvanath, G. Kim, and V. Govindaraju. Chaincode Contour Processing for Handwritten Word Recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, pages 928–932, 1999.

[14] M. Okamoto, S. Sakaguchi, and T. Suzuki. Segmentation of Touching Characters in Formulas. *Lecture Notes on Computer Science*, pages 151–156, 1999.

[15] Eric Lecolinet Richard G. Casey. A Survey of Methods and Strategies in Character Segmentation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 18(7):690–706, 1996.

[16] B.K. Sin and JH Kim. Ligature modeling for online cursive script recognition. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 19(6):623–633, 1997.

[17] X. Tian and Y. Zhang. Segmentation of Touching Characters in Mathematical Expressions Using Contour Feature Technique. In *Software Engineering, Artificial Intelligence, Networking, and Parallel/Distributed Computing, 2007. SNPD 2007. Eighth ACIS International Conference on*, volume 1, 2007.

[18] HM Twaakyondo and M. Okamoto. Structure analysis and recognition of mathematical expressions. In *Document Analysis and Recognition, 1995., Proceedings of the Third International Conference on*, volume 1, 1995.

[19] S. Uchida, A. Nomura, and M. Suzuki. Quantitative analysis of mathematical documents. *International Journal on Document Analysis and Recognition*, 7(4):211–218, 2005.

[20] H. Xue and V. Govindaraju. On the Dependence of Handwritten Word Recognizers on Lexicons. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, pages 1553–1564, 2002.

[21] R. Yamamoto, S. Sako, T. Nishimoto, and S. Sagayama. On-Line Recognition of Handwritten Mathematical Expressions Based on Stroke-Based Stochastic Context-Free Grammar. In *International Workshop on Frontiers in Handwriting Recognition. IEEE Computer Society*, pages 249–254, 2006.

Vita

Amit Arun Pillay was born in Mumbai, India on September 21, 1984, the son of Arun Pillay and Uma Pillay. He received the Bachelor of Engineering degree in Information Technology from Veermata Jijabai Technological Institute, Mumbai, India in 2006. He is currently pursuing his Master of Science degree from Rochester Institute of Technology, United States of America. His research interest includes Pattern Recognition, Computer Vision, Image Processing and Machine Learning. His current research includes combining syntactical pattern recognition techniques.

Permanent address: 1814 Crittenden Road Apt 6
Rochester, New York 14623

This thesis was typeset with LaTeX[†] by the author.

[†]LaTeX is a document preparation system developed by Leslie Lamport as a special version of Donald Knuth's TeX Program.