様式 C-7-2

自己評価報告書

平成 21 年 03 月 31 日現在

研究種目：基盤研究（Ｂ）
研究期間：平成１８年度～平成２１年度
課題番号：18300035
研究課題名（和文）　高精度数式認識と科学技術文書電子化システムの実用化のための研究開発
研究課題名（英文）　Research on High Accuracy Mathematical Formula Recognition and the Development of Digitization System of Scientific Documents.
研究代表者
鈴木　昌和（SUZUKI MASAKAZU）
九州大学・大学院数理学研究院・教授
研究者番号：20112302

研究分野：　数理情報システム学
科研費の分科・細目：情報学・メディア情報学・データベース
キーワード：数式認識、文字認識、科学文書電子化、電子ジャーナル、視覚障害者支援

１．研究計画の概要
本研究では、代表者らがこれまで行ってきた、数式を含む文書認識システムの実用化を目指して研究を更に進め、高い水準の科学技術文書電子化システムの実現を目指して研究を進める。具体的には、(1)複数の文字認識手法を組み合わせて Voting を行うことによる、個別文字・記号認識率の向上、(2)数式中の接触文字や分離文字に対応できる数式構造解析手法の研究、(3)認識実行時に数式を記述するパラメータを自動取得し、文書適応型の構造解析手法の導入、(4)数学文書のためのレイアウト解析と文書論理構造解析、表認識の精度向上のための研究を行う。

２．研究の進捗状況
(1) 認識精度を向上するために大規模な正解付き文字・数式画像データベースを構築し公開した(http://www.inftyproject.org/). 公開以来内外の研究者により今日まで途切れることなく毎週のようににダウンロードされている。そのデー手ベース上でSVMを用いた数学記号の詳細識別プログラムを開発した。また、動的計画法を用いて複数のＯＣＲエンジンを組み合わせる手法を導入し、現在はそれを数式のような２次元構造認識に拡張中である。
(2) 数式認識や文書論理構造解析に置いては文字の大きさや位置関係が重要であり、使用されているフォントや印刷スタイルによる影響は不可避である。そのため、認識対象とするドキュメントから自動的に各種のパラメータを取得し、文字認識や文書論理構造の精度を向上させる適応型の認識手法を開発した、実際に数学の論文誌やシリーズの書籍などの認識で大きな効果が得られることを確認した。
(3) 通常のテキスト認識では言語情報を用いることにより認識率を向上させる手法が有効であるが、数式認識には言語情報は利用できない。そのため、認識を評価する手法として、数式生成文法よる数式の「妥当性」を評価する仕組みの研究を進めた。この研究は一定の効果は確認できているものの、未だ研究途上の段階と言える。

３．現在までの達成度
大旨順調に進展している。
[理由] 当初計画していた主要な研究課題についてそれぞれ成果が得られている。また、研究成果はすべてプログラム実装も行っており、公開中の数式認識ソフトウェアに組み込み可能、または組み込み可能な形で成果が得られている。適応型認識については、実際の数学論文誌遡及電子化で有効性が確認され成果が生かされている。

４．今後の研究の推進方策
特に適応型認識において、論文誌のように均質性のある大量文書の場合には非常に大きな成果が上げられることが確認できているが、セミナー報告集などの多様な印刷スタイルの論文などが混在する場合には、まだ課題が多い。今後は数頁～数十頁単位の文書に対する適応型認識手法の開発に重点を置いて研究を進める。また、数式の生成文法を用いた「数式の妥当性」評価に慣用的構文や隣接記号出現頻度などの統計的データを組み込んだ評価法を導入する予定である。

５． 代表的な研究成果
（研究代表者、研究分担者及び連携研究者には下線）

〔雑誌論文〕（計 4 件）

① C.Malon, S. Uchida, M. Suzuki, Mathematical symbol recognition with support vector machines, Pattern Recognition Letters 29 (2008) 1326-1332, 査読有り
② .Aly, S.Uchida, M.Suzuki, Identifying Subscripts and Superscripts in Mathematical Documents, Mathematics in Computer Science, Vol.2, N.2, 195-209, 1998, 査読有り
③ A.Fujiyoshi, M.Suzuki, S.Uchida, Verification of Mathematical Formulae Based on a Combination of Context-Free Grammar and Tree Grammar, lecture Notes in Artificial Inteligence, Vol 5144, 2008, pp415-429, 査読有り
④ 金堀利洋、鈴木昌和 、PDF中のテキスト情報を利用した視覚障害者のための英文PDF科学技術文書読取りシステム、電子情報通信学会論文誌 D Vol.J90-D No.3 pp.706-714、査読有り

〔学会発表〕（計 件）

①「Multi-lingual mathematical document recognition by InftyReader, The 2nd @Science Thematic Network International Conference, Oct. 20th, 2008, Milan, Italy.
②「Multi-lingual Support in Infty Software, Adaptive Content Processing Conference Nov. 6, 2008, Amsterdam
③「A Large-Scale Analysis of Mathematical Expressions for an Accurate Understanding of Their Structure, The Eighth International Workshop on Document Analysis Systems, 2008.09, Nara, Japan
④「Accessing Mathematical print Material through the Infty System, International Conference ICCHP, July 8, 2008, Linz
⑤「Math-Document Accessibility with InftyReader and ChattyInfty, International Conference CSUN, March 13, 2008, Los Angeles.

〔その他〕