様式 C-19

科学研究費補助金研究成果報告書

平成　22　年　5　月　24　日現在

研究種目：基盤研究Ｂ（１）
研究期間：２００６～２００９
課題番号：１８３００００３５
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 Development of Digitization System of Scientific Documents
研究代表者
鈴木　昌和（SUZUKI MASAKAZU）
九州大学・数理学研究院・教授
研究者番号：20112302

研究成果の概要（和文）：科学技術文書のスキャン画像を検索や音声や点字などのアクセシブルなデータに変換可能な電子データに変換するシステム構築に不可欠な数式認識と、数式を含んだ文書のレイアウト解析の高精度化に関する研究を行った。特に大量の頁の文書の電子化に有効な適合型認識システムのアルゴリズムを文字認識、数式構造解析、レイアウト解析の各レベルで開発し実装を行った。また、類似記号が多い数式の文字認識精度向上のため、サポートベクターマシンを用いた類似数学記号識別の評価テストも行った。

研究成果の概要（英文）：This research is focused on the development of high performance system of OCR system for scientific documents. To improve the accuracy of mathematical formulae recognition and layout analysis of mathematical papers, we developed adaptive method efficient in large volume digitization. Distinction of similar mathematical symbols using support vector machine is also implemented and evaluated.

交付決定額

（金額単位：円）

	直接経費	間接経費	合　計
2006 年度	5,100,000	1,530,000	6,630,000
2007 年度	3,300,000	990,000	4,290,000
2008 年度	3,300,000	990,000	4,290,000
2009 年度	3,300,000	990,000	4,290,000
年度			
総　計			

研究分野：総合領域
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．研究開始当初の背景

科学技術文書のデータベース化は急速に本格化しつつあり、(図表や)数式を含んだ科学技術文書の認識・電子化システムの実用化を目指した研究の重要性は高まっている。

また、ＯＣＲを用いた印刷文書の自動点訳や読み上げは、目の不自由な人々の社会進出に大きな役割を果たしており、科学技術分野でも同様の環境を実現するために、数式を含む文書の認識は福祉の面から切実に求められている。

しかし、数式を含んだ科学技術文書の認識ができる商用ＯＣＲソフトは世界的に見ても未だない。一般的な WEB のアクセシビリティは年々向上しているが、科学的コンテンツは数式や図があり、PDF 内の数式は OCR

と数式認識の技術を用いなければ視覚障害者は読むことが出来ない。電子化された科学文書をアクセシブルにするためにも数式認識は必要となる。

２．研究の目的

本研究の動機が数学の学術文献の電子化や視覚障害者の数学的コンテンツへのアクセシビリティ支援に根ざしているため、単なる基礎研究に止めず、実際の電子化に有用なアプローチに焦点を合わせて研究を進めた。そして、本研究での到達目標を、数式を含む科学技術文書について、一般文書における既存商用ＯＣＲソフトと同水準以上の認識性能をもつ、実用ソフトウェアの開発においた。

代表者らが本研究の開始時点までに開発して来たシステム(以下、Infty と記す) ではテキスト領域に対する文字認識率は国産商用 OCR ソフトウェアとほぼ同水準に達していたが、数式領域では未だテキスト領域のほぼ倍の誤認識を含んでいた。本研究では、数式中の認識率をテキスト領域と同水準に向上させることを目標とした。

また、数式構造解析については、文字の誤認識や接触文字・分離文字に対して脆弱であったため、そうした低品質画像や、添え字の大きさなどが通常と異なる場合などに頑健なアルゴリズムの開発を目指した。

更に、学術雑誌の電子化や自動点訳などでは、タイトル頁の構造、章節等の構造、定義、定理、命題などの記述、数式番号や文献引用構造、文献表などの文書としての構造理解も非常に重要である。本研究では科学技術分野の論文や教科書・参考書などを中心に、自動解析できる範囲の拡大を目指した。また、そうした構造情報抽出の結果の誤りを容易に検出し修正するユーザーインターフェースの開発も同時に行った。

３．研究の方法

数式には多様な記号が使われ、類似記号が多い。そのため類似記号識別の為にサポートベクターマシンを用いた。また、文字認識の精度向上には文字切り分けの精度向上が重要であるが、数式では通常の文章の認識で使われる一次元の動的計画法が適用できない。また、bigram など言語情報に対応するコスト補正による精度向上が出来ない。そこで、大量の文書認識に有効な適合型のアルゴリズムを導入した。認識対象とする文書画像から文字特徴や添え字などの大きさ・号配置に関する特徴を抽出し、文字認識と構造解析の精度を高めるアプローチを取った。タイトル、章、節、定理記述などの論理構造解析にも同様の適合型のアルゴリズムを採用した。

OCR を用いる処理では誤認識は避けて通れない問題である。自動認識による精度の向上を図る一方で、文字・数式の認識や論理構造解析の認識の誤りを検出し修正するユーザーインターフェースの開発も変更して行った。逐次結果を修正するインターフェースだけでなく、認識対象とする文書に含まれる文字や記号の特徴を抽出して編集して認識処理に反映させるユーザーインターフェースの開発も行った（後述の適合型処理を参照)。

４．研究成果

１．データベース

初年度はその後の研究の基盤となるデータベース整備を行った。

文字認識と数式構造解析の正解情報(GroundTruth) を付与した文字・数式画像データベース構築し、InftyCDB-1 ～ InftyCDB-3 と３セットにして公開した (http://www.inftyproject.org/)。

このデータベースは全体で約 150 万文字を含む大規模なものであり、国際的にも数式を含む GroundTruth 付きデータベースとしては他に類を見ない最大規模のものである。国際会議で配布した数と、内外の研究者による WEB からのダウンロード数を合わせると３００を超える。

また、著作権法上の問題があり、公開はしていないが、数学の論文誌や単行本の文書画像に文字・数式構造情報だけでなく文書論理構造のタグも含めた詳細な Ground Truth のデータベースも作成し、文書論理構造解析の研究に用いた。

２．文字・数学記号認識

（１）数式中の類似文字識別

数学専門書で使われる数学記号の種類は非常に多く、類似した記号も多くある。また、単に文字コードだけでなく、数式ではフォントにより異なる概念のものを表すことが通例であり、単語単位ではなく孤立して用いられる数学記号のフォント識別も重要な課題である。そのため、サポートベクターマシンを用いた類似記号の詳細識別器開発とその評価実験を行った。サポートベクターマシンは原理的に２クラス識別機であるが、従来の認識エンジンと Confusion Matrix を用いて組み合わせることにより、効果的な詳細識別器を構成することが可能であり、認識率の向上に大きく寄与することを実験により確認した。特に、フォント識別能力が非常に高く、数式中に現れるスクリプト体文字やドイツ文字、類字ラテン文字をもつギリシャ文字の識別などに有効であることを確認した。

（２）数式領域の抽出

通常のテキスト領域と数式領域では文字や記号の切り分け（セグメンテーション）手法や利用可能な誤認識補正処理の原理が異

なる。そのため、全体の認識率向上のためにはテキスト領域と数理器領域の切り分けが重要になる。

本研究では高精度の認識システムを開発するために複数の認識手法を組み合わせる方法でテキスト領域の文字認識精度の向上を計った。独自に開発している認識エンジンの認識結果の信頼度補正プログラムを加え、更に異なるメーカーで開発された２つ認識エンジンを用いて、動的計画法により局所的最適解の列を取得する方法が数式の前後での通常文字の認識率向上に非常に有効であることが判明した。メーカーの認識エンジンは数式部を通常文字として認識し、しかもその前後で異なった誤認識を発生させることが多い。そのため、独自の数学記号を認識出来るOCRエンジンに加えて、複数のメーカーによるOCRエンジンを補助的に動的計画法のコストに反映させることにより、一層正確な数力領域切り分けが可能になった。

３.数式構造解析

（１）パラメータ抽出

数式構造解析では隣接文字の接続関係（水平、上付き添え字、下付添え字など）の判定が重要である。その為、実際の文書画像の正解付きデータベースから記号のカテゴリー対毎に添え字とベースライン文字の大きさの比と正規化した矩形中心座標の相対位置の分布図を作成して判定のコスト関数導いている。このコスト関数の導出には各種のパラメータを用いているが、従来は前データベースから算出した統計データにより、包括的に与えていた。本研究では数式構造解析の精度向上を計るため、文書毎に異なる数式記述の各種パラメータの自動的に事前推定してコスト関数を導出する適合型の数式認識アルゴリズムを開発し、数式構造解析の高精度化に有効であることを実験により確認した。

（２）接触文字と分離文字

数式中に接触文字や分離文字が含まれていても、複数の文字切り分け候補の中から正しい切り分け結果を選択する評価関数に、文字認識のスコアだけでなく隣接文字の大きさと位置関係評価値を加え、精度向上を図った。上述のコスト関数の補正は添え字判定等の構造解析に有効なだけでなく、文字切り分けのコストにも反映することにより、文字認識精度向上にも繋がるものである。

（３）文法を用いた構造解析

高速なパーシングを可能とする Linear Monaic Tree 文法による新たに数式生成文法を導入し、数式認識結果の評価と誤認識場所を発見する関数をつくり、実証実験を行った。

（４）行列の認識については隣接行列要素との境目の判定に数式構文構造を加味し、精度向上をはかった。

４．大量文書に対する適合型処理

本研究を進める過程で、代表者らはNPO法人サイエンス・アクセシビリティ・ネットと協力して実際の数学論文誌の遡及電子化作業に研究成果を反映させた。日本数学会の数学論文誌（J. Math. Soc. Japan）を始め、いくつかの国内の英文数学論文誌や数理解析研究所の講究録など、合計約 20 万頁に及ぶ電子化作業の中で研究成果を反映させ、またそこで発生した問題を研究にフィードバックする形で研究開発を進めた。認識処理した結果はPDF形式でスキャン画像の背後に検索用のテキスト文字と数式構造解析結果をLaTeX ソースの形式で埋め込み、更にいくつかの論文誌では、各論文のセクション等のしおり、定理命題へのリンク、文献引用記述部から文献表へのリンクなどを埋め込んだ PDF を生成した。その結果は Project Euclid などの実際に公開されているオンラインジャーナルで見ることが出来る。

（１）文字・記号認識

大量の文献を電子化する際に認識対象とする文書画像から自動的に文字抽出してカテゴリー毎にクラスタリングして、認識辞書に登録する手法による認識率向上を行った。

同時に、文書中の文字画像を分類して抽出し、画面に一覧して大量の文字のクラスタリング結果や認識結果を短時間で確認・修正できるユーザーインターフェースを開発し、実際に数学の学術論文誌の電子化などの作業でその著しい有効性を確認した。この方法は、テキスト領域中の文字だけでなく数式中の文字や記号の認識率向上にも著しい効果があった。

（２）文書論理構造解析

認識対象とする書籍や論文誌から、行単位の文字の大きさや太さ、書体などの特徴量を抽出し、それを自動分類する実験を行った。また、その行特徴量を用いた大小関係を定義し、章、節、小見出しの判定、行特徴量とキーワードを組み合わせ、定理、補題、系などの命題記述部の抽出を行い、実際の数学論文誌電子化でその効果を実証した。

５．主な発表論文等

（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 5 件）

1. T.Kanahori, A.Sexton, V.Sorge, M.Suzuki, Capturing Abstract Matrices from Paper, Lecture Notes in Computer Science 4108 (2006) pp.124-138

2. C.Malon, S.Uchida, M.Suzuki, Mathematical symbol recognition with support vector machines, Pattern

Recognition Letters, vol.29, no.9 (2008) pp.1326-1332
3. W. Aly, S. Uchida, M. Suzuki, Identifying Subscripts and Superscripts in Mathematical Documents, Mathematics in Computer Science, Vol.2, No 2, 2008, 195-2091
4. W. Aly, S. Uchida, M. Suzuki, Automatic Classification of Spatial Relationships among Mathematical Symbols Using Geometric Features, IEICE Transactions on Information & Systems, vol.E92-D, no.11 pp.2235-2243, 2009.
5. A. Fujiyoshi, M. Suzuki, S. Uchida, Grammatical Verification for Mathematical Formula Recognition Based on Context-Free Tree Grammar, Mathematics in Computer Science, Vol. 3, No. 3, 2010, pp.279-298

〔学会発表〕(計８件)
1. T.Kanahori, M.Suzuki, Refinement of retro- digitized documents including mathematical formulae by re-recognition, 2nd IEEE International Conference on Document Image Analysis for Libraries, April 27-28, 2006, Lyon, France.
2. Handwriting Interface to Mathematical Expressions, East Coast Computer Algebra Day, May, 2006, Drexel University, Philadelphia
3. M.Suzuki, A Handwriting Interface for Mathematical Expressions, Invited Session Pen-Based Interfaces for Mathematical Documents, Communicating Mathematics in Digital Era, CMDE 2006 August 15-18, 2006, Aveiro, Portugal
4. M.Suzuki, Multi-lingual mathematical document recognition by InftyReader, the 2nd @Science Thematic Network International Conference, October 20, 2008, Milan, Italy.
5. M.Suzuki, Virtual Link Network Method to recognize Printed Mathematical Formulae, Seminar in Informatics, Universit`a degli Studi di Milano, October 21, 2008, Milan,Italy.
6. W. Aly, S. Uchida, and M. Suzuki, A Large-Scale Analysis of Mathematical Expressions for an Accurate Understanding of Their Structure, The 8th International Workshop on Document Analysis Systems, September, 2008, Nara, Japan
7. W. Aly, S. Uchida, A. Fujiyoshi, M. Suzuki, Statistical classification of spatial relationships among mathematical symbols, The 10th International Conference on Document Analysis and Recognition, July, 2009, Barcelona, Spain
8. A. Fujiyoshi, M. Suzuki, S. Uchida, Syntactic Detection and Correction of Misrecognitions in Mathematical OCR The 10th International Conference on Document Analysis and Recognition, July, ICDAR 2009, Barcelona, Spain

〔図書〕(計０件)
〔産業財産権〕
○出願状況 (計０件)

〔その他〕
URL: http://www.inftyproject.org

６．研究組織
(1)研究代表者
鈴木　昌和 (SUZUKI MASAKAZU)
九州大学数理学研究院・教授
研究者番号：20112302

(2)研究分担者
内田　誠一 (UCHIDA SEIICHI)
九州大学・ステム情報学研究院・教授
研究者番号：70315175
(H20→H21：連携研究者)
岡本　正行 (OKAMOTO MASAYUKI)
信州大学・工学部・教授
研究者番号：50109196
(H20→H21：連携研究者)
玉利　文和 (TAMARI FUMIKAZU)
福岡教育大学・教育学部・教授
研究者番号：70036937
(H20→H21：連携研究者)
藤本　光史 (FUJIMOTO MITSUSHI)
福岡教育大学・教育学部・准教授
研究者番号：20270241
(H20→H21：連携研究者)
金堀　利洋 (KANAHORI TOSHIHIRO)
筑波技術大学・障害者高等教育センター・准教授
研究者番号：00352568
(H20→H21：連携研究者)
山口雄仁 (YAMAGUCHIKATSUHITO)
日本大学短期大学部・教授
研究者番号：00182428

(3)連携研究者
藤芳　明生
茨城大学・工学部・講師
研究者番号：00323212