様式Ｃ－１９

科学研究費助成事業（科学研究費補助金）研究成果報告書

平成　２４年　６月　１９日現在

機関番号：13601
研究種目：基盤研究(C)
研究期間：2009　～　2011
課題番号：21500163
研究課題名（和文）文書解析・認識オープンソース OCRopus への数式認識モジュールの組込み

研究課題名（英文）Embedding a Mathematical OCR Module into OCRopus

研究代表者
岡本正行（　OKAMOTO MASAYUKI　）
信州大学工学部・教授
研究者番号：50109196

研究成果の概要（和文）：
本研究は学術雑誌等の科学技術文献を対象として，数式を含めた全文を読取る OCR（光学的文字読取り装置）の開発に寄与することを目的とし，ドイツ人工知能研究所(DFKI)が開発を行なっているオープンソースソフト OCRopus への数式認識モジュールの組込みを行った．組込みに際しては，組込み位置の検討を行い，既存モジュールの機能を活用することでシステムとの親和性を高めた．数式が単独の行で出現するディスプレイ数式については，自動的に数式位置の判別および切出しを行い，モジュール化した数式認識により Math-ML や LaTex での出力を得ることに成功した．

研究成果の概要（英文）：
In this research, we aim to contribute the development of the full text OCR system that is also able to read mathematical expression, and embed our mathematical OCR module to an open source OCR soft OCRopus developed by DFKI (German Research Center for Artificial Intelligence). To increase the affinity between our module and the system, we reviewed the embedding position of our module and utilized the existence functions of the OCRopus. As for the display expression in which the mathematical expression appears as a line independently, our module is able to locate and crop the line automatically, then the cropped line is correctly converted to the Math-ML or the LaTex format.

交付決定額

（金額単位：円）

	直接経費	間接経費	合　計
2009 年度	1,500,000	450,000	1,950,000
2010 年度	900,000	270,000	1,170,000
2011 年度	900,000	270,000	1,170,000
年度			
年度			
総　計	3,300,000	990,000	4,290,000

研究分野：画像処理・パターン認識
科研費の分科・細目：知能情報処理・知能ロボティクス
キーワード：パターン認識

１．研究開始当初の背景

印刷文書を電子文書に変換する自動的な処理（OCR：Optical Character Recognition）は，「自炊」という用語の普及に見られるように近年エンドユーザに対して幅広く普及し関心が高まっている．その要因の一つとして，電子書籍リーダの普及やデータのクラウド化により，電子データの携帯性が向上し需要が高まったことや，電子ジャーナルや電子図書館の登場により，書籍を電子データとして入手するシステムも整備されてきたことがある．

文書認識・解析の研究は，1990 年代より従来の単純な OCR 機能を拡張するために精力的に行われてきた．主な処理としては，スキャンした文書画像の傾き検出・補正，レイアウト解析，論理構造解析，文字の切出しおよび認識があり，各々について研究が行われている．その結果多様な印刷品質，レイアウト，言語等に対応した OCR 製品も実現されてきているが，未だに様々な印刷文書を電子化するためには克服すべき課題が山積している．この分野の研究を行うためにシステムを試作して性能評価を行う場合には，上述した一連の処理を自分で実装する必要があるが，この作業は容易なことではない．このためドイツ人工知能研究所の IUPR グループでは Google より資金援助を受け，文書解析・認識研究のためのモジュール OCRopus をオープンソースとして公開し，この分野の研究促進を図っている．

申請者は平成 9 年以来，ドイツエッセン大学実験数学研究所の Michler 教授により開始された「数学文献のディジタル化プロジェクト」において数式認識エンジンの開発を担当し，その推進に寄与してきた．以降このプロジェクトは九州大学鈴木昌和教授らが開発してきた InftyReader に受け継がれてきたが，今後もさらに高性能な科学技術文献用 OCR を開発していくためには，OCRopus の一モジュールとして数式認識エンジンを公開し，この分野の研究を促進させることが重要となる．

２．研究の目的

本研究では OCRopus 用の数式認識モジュールの開発を目的とする．OCRopus が読取り対象とする文書は科学技術系の論文や専門書を含むが，文書が数式を含む場合，数式が単語として誤認識され，以降の文書の読取に影響を及ぼす．これを回避するためには，数式部分を把握しテキスト文書と数式で処理を分ける必要がある．一方，申請者らの研究グループは，前章で述べたように，数式認識用のシステムを開発してきた．ここでは独自の前処理，レイアウト解析，OCR 等を実装し実験を行なってきたが，特定の文書スタイルに特化されており汎用性に乏しかった．以上の理由より，OCRopus に当研究室で開発した数式認識モジュールを組込むことにより，数式認識機能を持つより汎用的なシステムの構築を目指している．

OCRopus を全体的な OCR システムとして用いる理由は以下のとおりである．

- オープンソースのためソースコードが公開されており，独自の改良・改変が容易である
- 前処理，レイアウト解析，文字認識といった OCR に関わる処理を独立に実行可能で，各々の機能の性能評価を個別に行える
- 各モジュールのインターフェース仕様が明確であり，内蔵のスクリプト言語を用いれるなど拡張性が高く，プログラミングが容易である

OCRopus への数式認識モジュールの組込みにあたっては，当面は文書中のディスプレイ数式（後述 3.1 参照）のみを認識対象としている．

OCRopus の現状の性能，本研究室の有する数式認識の精度を鑑みて，本研究の掲げる達成目標を以下の 2 つとする．

- 通常のテキストラインとディスプレイ数式の識別法を提案する
- ディスプレイ数式に対しては数式認識を行うシステムを構築する

実験では，数式を含む文書画像全体の認識を行い，その認識結果を示す．

３．研究の方法

数式を含む文書全体の文字・数式認識を行うにあたり，使用した数式用 OCR について説明した後，全体の処理過程と，識別処理に使用したツールや手法について述べる．

(1) 数式認識モジュール

当研究室でこれまでに開発されてきた印刷数式認識システムの概要を紹介した後，OCRopus に数式認識モジュールを組込むための手順について述べる．数式認識システムは記号認識部と数式構造解析部の 2 つから構成されている．

① 記号認識部では，分離記号の統合，接触記号の分離を行った後，各記号の認識を行っている．

② 構造解析部は各記号の種類，サイズ，数式内での位置情報等を用いて数式構造を解析し，解析結果を LaTeX, MathML で出力す

る．

数式は，他のテキストラインとは独立に印刷されるディスプレイ数式と，通常のテキストラインの中に埋め込まれるインライン数式に区別される．

(2) 全体のフロー

数式認識モジュールを組込んだ全体の処理フローを図１に示す．黒塗り部は独自に実装した処理で，それ以外は OCRopus の機能を用いている．

① 数式を含んだ文書画像を入力する．
② 画像に対し，2 値化や傾き補正，ノイズ除去などの前処理を行う．
③ レイアウト解析を行い，文章領域とその他を判別し，文章領域については行毎のテキストラインに分割する．ここで，各テキストラインの外接矩形情報(座標値)を得る．
④ それぞれのテキストラインに対し文字分割を行う．ここで，記号毎の外接矩形情報(座標値)を得る．
⑤ テキストライン分割と文字分割によって得られた矩形情報により，テキストライン 1 行毎に特徴量を計算し，ディスプレイ数式と通常のテキストラインを識別する．
⑥ 数式であると識別されたテキストラインは数式認識モジュールに入力され，認識結果を出力する．
⑦ 非数式であると識別されたテキストラインは OCRopus により文字認識が実行される．
⑧ 両者の認識結果を統合する．
⑨ 文書画像全体の認識結果を出力する．

(3) レイアウト解析によるテキストライン分割

ディスプレイ数式の識別に用いる特徴は，テキストラインの外接矩形が適切に切り出されていないと，正しく計算出来ない．しかし，デフォルトの OCRopus のレイアウト解析によるテキストライン分割では，図２のようにテキストラインが正しく切り出されない場合がある．

そこで，テキストライン分割のアルゴリズムを検討する必要が出てきた．OCRopus ではレイアウト解析に用いるアルゴリズムを自由に選択できる仕様になっており，様々なアルゴリズムを独立に試すことが可能である．本研究で対象とした文書は数式を含む１段組のレイアウトであり，ここではテキストラインの切出しが最も良好なものは，1CP と呼ばれるアルゴリズムであった．

1CP は１段組の文書画像を入力に想定したレイアウト解析のアルゴリズムで，水平方向の周辺分布によって文章領域をテキストラインに分割している．

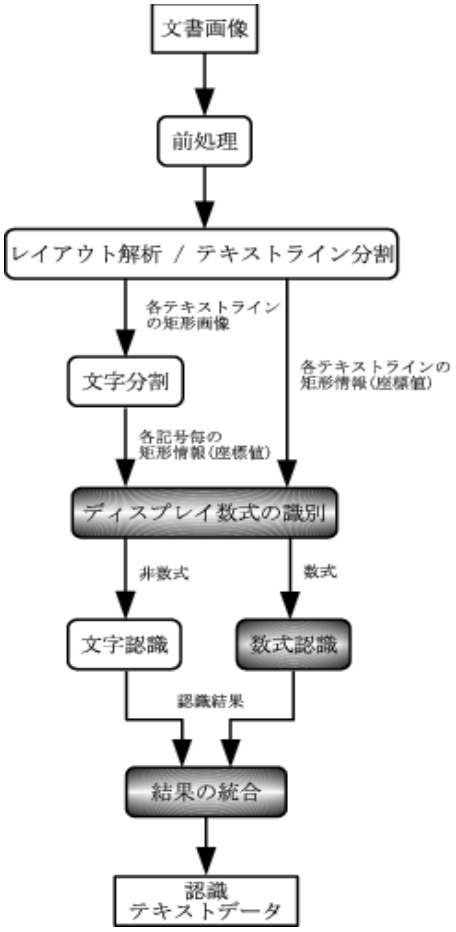
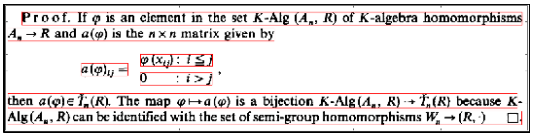


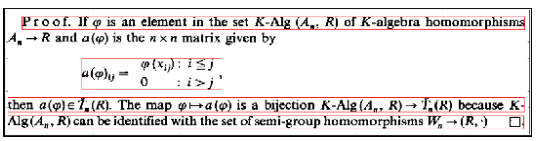
図 1 数式認識を含む OCR 処理全体のフローチャート．黒色処理部分が数式認識のために新たに追加した処理．


Proof. If φ is an element in the set K-Alg (A_n, R) of K-algebra homomorphisms $A_n \to R$ and $a(\varphi)$ is the $n \times n$ matrix given by

$$a(\varphi)_{ij} = \begin{cases} \varphi(x_{ij}) & : i \leq j \\ 0 & : i > j \end{cases},$$

then $a(\varphi) \in \bar{T}_n(R)$. The map $\varphi \mapsto a(\varphi)$ is a bijection K-Alg $(A_n, R) \to \bar{T}_n(R)$ because K-Alg (A_n, R) can be identified with the set of semi-group homomorphisms $W_n \to (R, \cdot)$ □

(a)デフォルトのテキストライン分割


Proof. If φ is an element in the set K-Alg (A_n, R) of K-algebra homomorphisms $A_n \to R$ and $a(\varphi)$ is the $n \times n$ matrix given by

$$a(\varphi)_{ij} = \begin{cases} \varphi(x_{ij}) & : i \leq j \\ 0 & : i > j \end{cases},$$

then $a(\varphi) \in \bar{T}_n(R)$. The map $\varphi \mapsto a(\varphi)$ is a bijection K-Alg $(A_n, R) \to \bar{T}_n(R)$ because K-Alg (A_n, R) can be identified with the set of semi-group homomorphisms $W_n \to (R, \cdot)$ □

(b) 1CP によるテキストライン分割

図 2 テキストライン分割の例

(4) 文字分割

OCRopus の文字分割機能により，それぞれのテキストライン矩形画像に対し文字分割を行い，文字毎の外接矩形情報を得る．図３

はディスプレイ数式とそうでないテキストライン画像に対し文字分割を行った結果である．

(a)ディスプレイ数式ではないテキストライン

(b)ディスプレイ数式のテキストライン

図 3　文字分割の例

(5) ディスプレイ数式の識別

① 数式の特徴

ディスプレイ数式の識別には．Garain 等の提案した特徴量（f_1 ．f ．f_{nh}）の他に，新たに提案したもの（f_{ma}, f_{mi}）を用いている．

$$f_{ws} = \frac{r}{r_\mu}, f_{ms} = \sigma_y, f_{mh} = \frac{h}{h_\mu}$$

ここで，r はテキストラインの上下の空白の平均，r_μ はすべての連続するテキストライン間の空白の平均，σ_y はテキストライン中にある記号の外接矩形の右下 y 座標の標準偏差，h はテキストラインの高さ，h_μ はすべての h の平均を表す．

新たに定義した特徴量は以下のとおりである．

$$f_{ma} = \sigma_a, f_{mi} = ind$$

ただし，σ_a はテキストライン中にある記号の外接矩形アスペクト比の標準偏差，ind はテキストラインの左インデントの距離である．

② 識別アルゴリズム

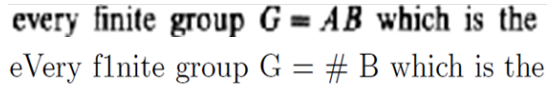
識別に関しては，Garain 等は上記の複数の特徴量を１つの値に統合し閾値処理を行っていたが，本研究では SVM を用いている．SVM には解析用ソフトウェアとして SVM-Light を用いている．

③ 学習

上記のフローと同じ手順で，入力文書画像に対し，テキストライン分割と文字分割を行い，特徴量を計算する．その後，手動でテキストラインに対してディスプレイ数式かどうかのラベルを付与し Ground Truth とする．この Ground Truth を用いて，SVM-Light の学習を行い識別器を作成した．

(6) 文字認識

上記識別処理で，数式とされなかったテキストラインに関しては，OCRopus の文字認識を行う．図４は非数式と識別されたテキストラインに対し，OCRopus の文字認識を実行した結果の例である．ここで，認識結果中の文字’#’ は認識できなかった文字を示している．



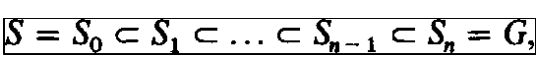
eVery f1nite group G = # B which is the

図 4　文字分割の例．

入力画像（上）と認識結果（下）

(7) 数式認識

識別処理で，数式とされたテキストラインに関しては，数式認識を行う．図５は数式と識別されたテキストラインに対し，数式認識モジュールを行った結果の例である．



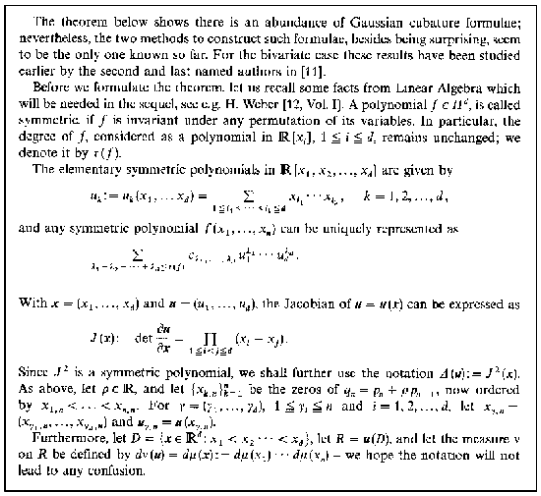

```
\begin{math}
S = {S }_{0 }^{}\subset {S }_{1 }^{}\subset
\dots \subset {S }_{n - 1 }^{}\subset {S }_{n }^{}= G
\end{math}
```

図 5　数式認識モジュールによる認識結果

入力画像（上）と LaTeX 変換結果（下）

４．研究成果

数式を含む文書画像に対し実験を行った．対象となる文書は１段組，図や表を含まないものであり 600dpi でスキャンした．図６はテストに用いた画像の例である．



The theorem below shows there is an abundance of Gaussian cubature formulae; nevertheless, the two methods to construct such formulae, besides being surprising, seem to be the only one known so far. For the bivariate case these results have been studied earlier by the second and last named authors in [11].

Before we formulate the theorem, let us recall some facts from Linear Algebra which will be needed in the sequel, see e.g. H. Weber [12, Vol. I]. A polynomial $f \in \Pi^d$, is called symmetric if f is invariant under any permutation of its variables. In particular, the degree of f, considered as a polynomial in $\mathbb{R}[x_i]$, $1 \leqq i \leqq d$, remains unchanged; we denote it by $r(f)$.

The elementary symmetric polynomials in $\mathbb{R}[x_1, x_2, \ldots, x_d]$ are given by

$$u_k := u_k(x_1, \ldots x_d) = \sum_{1 \leqq i_1 < \cdots < i_k \leqq d} x_{i_1} \cdots x_{i_k}, \quad k = 1, 2, \ldots, d,$$

and any symmetric polynomial $f(x_1, \ldots, x_n)$ can be uniquely represented as

$$\sum_{s_1 = i_1 - \cdots + i_d s = n f} c_{i_1, \ldots i_d} u_1^{i_1} \cdots u_d^{i_d}.$$

With $x = (x_1, \ldots, x_d)$ and $u = (u_1, \ldots, u_d)$, the Jacobian of $u = u(x)$ can be expressed as

$$J(x): \quad \det \frac{\partial u}{\partial x} = \prod_{1 \leqq i < j \leqq d} (x_i - x_j).$$

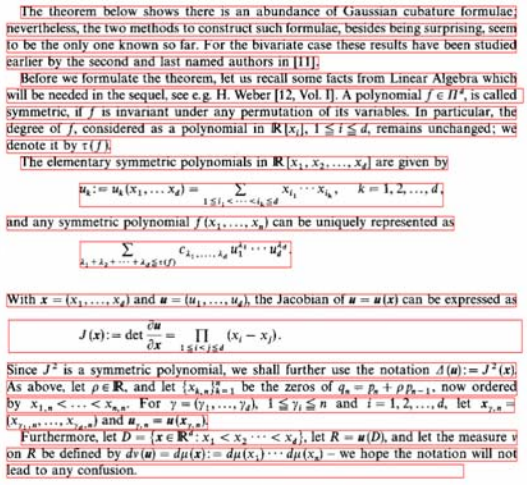
Since J^2 is a symmetric polynomial, we shall further use the notation $\Delta(u) := J^2(x)$. As above, let $\rho \in \mathbb{R}$, and let $\{x_{k,n}\}_{k=1}^n$ be the zeros of $q_n = p_n + \rho p_{n-1}$, now ordered by $x_{1,n} < \cdots < x_{n,n}$. For $\gamma = (\gamma_1, \ldots, \gamma_d)$, $1 \leqq \gamma_i \leqq n$ and $i = 1, 2, \ldots, d$, let $x_{\gamma,n} = (x_{\gamma_1,n}, \ldots, x_{\gamma_d,n})$ and $u_{\gamma,n} = u(x_{\gamma,n})$.

Furthermore, let $D = \{x \in \mathbb{R}^d : x_1 < x_2 \cdots < x_d\}$, let $R = u(D)$, and let the measure ν on R be defined by $d\nu(u) = d\mu(x) := d\mu(x_1) \cdots d\mu(x_n)$ – we hope the notation will not lead to any confusion.

図 6　テスト画像

(1) ディスプレイ数式の識別結果

プロジェクト”Retro-digitalization of mathematical journals, and their integration searchable digital libraries”で使われた数学系ジャーナルである”Archiv der Mathematik”をスキャンした画像データセット(600dpi, 1段組)に対し実験を行った．文書に図や表は含まれない．訓練用データ(正解データ)は 50 ページ，実験用は訓練用とは別の 64 ページから成る．訓練用は 1074 行の数式行と 174 行の非数式行を含む．実験用画像の一例を以下に示す．テキストライン分割を行い(図 7 (a))，ディスプレイ数式とそれ以外のテキストラインの識別を行った結果(図 7 (b))の図 8 (b)の黒塗り部がディスプレイと識別されたテキストラインである．


The theorem below shows there is an abundance of Gaussian cubature formulae; nevertheless, the two methods to construct such formulae, besides being surprising, seem to be the only one known so far. For the bivariate case these results have been studied earlier by the second and last named authors in [11].

Before we formulate the theorem, let us recall some facts from Linear Algebra which will be needed in the sequel, see e.g. H. Weber [12, Vol. I]. A polynomial $f \in \Pi^d$, is called symmetric, if f is invariant under any permutation of its variables. In particular, the degree of f, considered as a polynomial in $\mathbb{R}[x_i]$, $1 \leqq i \leqq d$, remains unchanged; we denote it by $\tau(f)$.

The elementary symmetric polynomials in $\mathbb{R}[x_1, x_2, \ldots, x_d]$ are given by

$$u_k := u_k(x_1, \ldots x_d) = \sum_{1 \leqq i_1 < \cdots < i_k \leqq d} x_{i_1} \cdots x_{i_k}, \quad k = 1, 2, \ldots, d,$$

and any symmetric polynomial $f(x_1, \ldots, x_n)$ can be uniquely represented as

$$\sum_{\lambda_1 + \lambda_2 + \cdots + \lambda_d \leqq \tau(f)} c_{\lambda_1, \ldots, \lambda_d} u_1^{\lambda_1} \cdots u_d^{\lambda_d}.$$

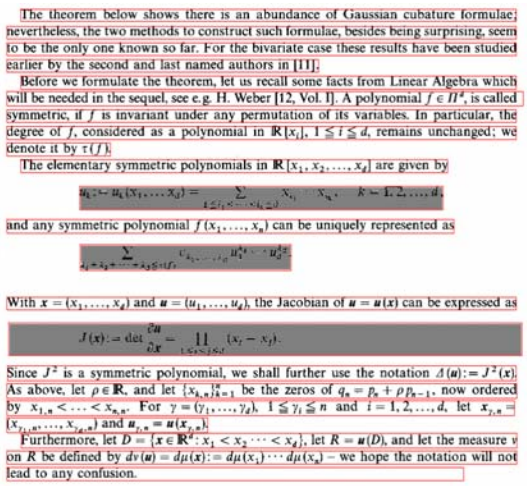
With $\boldsymbol{x} = (x_1, \ldots, x_d)$ and $\boldsymbol{u} = (u_1, \ldots, u_d)$, the Jacobian of $\boldsymbol{u} = \boldsymbol{u}(\boldsymbol{x})$ can be expressed as

$$J(\boldsymbol{x}) := \det \frac{\partial \boldsymbol{u}}{\partial \boldsymbol{x}} = \prod_{1 \leqq i < j \leqq d} (x_i - x_j).$$

Since J^2 is a symmetric polynomial, we shall further use the notation $\Delta(\boldsymbol{u}) := J^2(\boldsymbol{x})$. As above, let $\rho \in \mathbb{R}$, and let $\{x_{k,n}\}_{k=1}^n$ be the zeros of $q_n = p_n + \rho p_{n-1}$, now ordered by $x_{1,n} < \cdots < x_{n,n}$. For $\gamma = (\gamma_1, \ldots, \gamma_d)$, $1 \leqq \gamma_i \leqq n$ and $i = 1, 2, \ldots, d$, let $\boldsymbol{x}_{\gamma,n} = (x_{\gamma_1,n}, \ldots, x_{\gamma_d,n})$ and $\boldsymbol{u}_{\gamma,n} = \boldsymbol{u}(\boldsymbol{x}_{\gamma,n})$.

Furthermore, let $D = \{\boldsymbol{x} \in \mathbb{R}^d : x_1 < x_2 \cdots < x_d\}$, let $R = \boldsymbol{u}(D)$, and let the measure ν on R be defined by $d\nu(\boldsymbol{u}) = d\mu(\boldsymbol{x}) := d\mu(x_1) \cdots d\mu(x_n)$ – we hope the notation will not lead to any confusion.

(a)


The theorem below shows there is an abundance of Gaussian cubature formulae; nevertheless, the two methods to construct such formulae, besides being surprising, seem to be the only one known so far. For the bivariate case these results have been studied earlier by the second and last named authors in [11].

Before we formulate the theorem, let us recall some facts from Linear Algebra which will be needed in the sequel, see e.g. H. Weber [12, Vol. I]. A polynomial $f \in \Pi^d$, is called symmetric, if f is invariant under any permutation of its variables. In particular, the degree of f, considered as a polynomial in $\mathbb{R}[x_i]$, $1 \leqq i \leqq d$, remains unchanged; we denote it by $\tau(f)$.

The elementary symmetric polynomials in $\mathbb{R}[x_1, x_2, \ldots, x_d]$ are given by

and any symmetric polynomial $f(x_1, \ldots, x_n)$ can be uniquely represented as

With $\boldsymbol{x} = (x_1, \ldots, x_d)$ and $\boldsymbol{u} = (u_1, \ldots, u_d)$, the Jacobian of $\boldsymbol{u} = \boldsymbol{u}(\boldsymbol{x})$ can be expressed as

Since J^2 is a symmetric polynomial, we shall further use the notation $\Delta(\boldsymbol{u}) := J^2(\boldsymbol{x})$. As above, let $\rho \in \mathbb{R}$, and let $\{x_{k,n}\}_{k=1}^n$ be the zeros of $q_n = p_n + \rho p_{n-1}$, now ordered by $x_{1,n} < \cdots < x_{n,n}$. For $\gamma = (\gamma_1, \ldots, \gamma_d)$, $1 \leqq \gamma_i \leqq n$ and $i = 1, 2, \ldots, d$, let $\boldsymbol{x}_{\gamma,n} = (x_{\gamma_1,n}, \ldots, x_{\gamma_d,n})$ and $\boldsymbol{u}_{\gamma,n} = \boldsymbol{u}(\boldsymbol{x}_{\gamma,n})$.

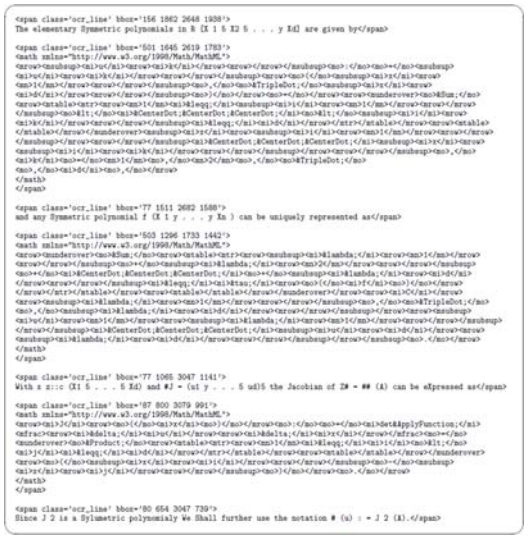
Furthermore, let $D = \{\boldsymbol{x} \in \mathbb{R}^d : x_1 < x_2 \cdots < x_d\}$, let $R = \boldsymbol{u}(D)$, and let the measure ν on R be defined by $d\nu(\boldsymbol{u}) = d\mu(\boldsymbol{x}) := d\mu(x_1) \cdots d\mu(x_n)$ – we hope the notation will not lead to any confusion.

(b)

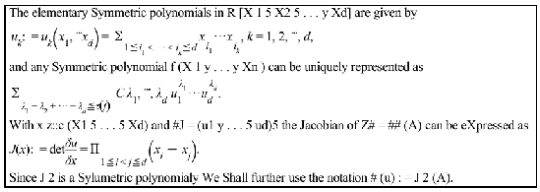
図 7　テキストライン分割 (a) とディスプレイ数式の識別結果 (b)

(2) 最終的な認識結果

ディスプレイ数式の数式認識結果と，OCRopus で出力されたテキストラインの認識結果を統合した結果の一部を図 8 に示す．数式認識の結果は MathML 形式，テキストラインの認識結果は HTML を用いて表現しており，両者の認識結果を統合することは容易である．



(a) HTML ソースコード


The elementary Symmetric polynomials in R [X 1 5 X2 5 . . . y Xd] are given by

$$u_k := u_k\left(x_1, \ldots x_d\right) = \sum_{1 \leq i_1 < \cdots < i_k \leq d} x_{i_1} \cdots x_{i_k}, k = 1, 2, \ldots, d,$$

and any Symmetric polynomial f (X 1 y . . . y Xn) can be uniquely represented as

$$\sum_{\lambda_1 + \lambda_2 + \cdots + \lambda_d \leq \tau(f)} C\lambda_1, \ldots, \lambda_d u_1^{\lambda_1} \cdots u_d^{\lambda_d}.$$

With x z::e (X1 5 . . . 5 Xd) and #J = (u1 y . . . 5 ud)5 the Jacobian of Z# = ## (A) can be eXpressed as

$$J(x): = \det \frac{\partial u}{\partial x} = \prod_{1 \leq i < j \leq d} \left(x_i - x_j\right).$$

Since J 2 is a Sylumetric polynomialy We Shall further use the notation # (u) : = J 2 (A).

(b) 認識結果表示

図 8　最終的な認識結果

(3) 考察

① 数式の検出

本研究で用いたデータセットは数式系のジャーナルであるため，ディスプレイ数式に限らずインライン数式も多く含まれる．したがって，提案した特徴量が高くなる問題がある．逆に，記号を囲む矩形の底辺が垂直方向で大きな変化が見られない場合，特徴量が低くなり数式が見落とされることがある．

② レイアウト解析の問題

レイアウト解析のアルゴリズムを 1CP に変更したことにより，多段組の文書を処理することができない．また，テキストラインの外接矩形が互いに重なり合っていると一つの行として分割されてしまう問題がある．また，次のような数式：

$$\sum_{i=1}^{n} i = n(n+1)/2$$

は複数行に分割される．これは，水平方向の周辺分布を用いているためである．

③ 文字認識の問題点
2012 年 6 月現在で OCRopus は ASCII 文字でも記号を正しく認識できない．その理由は以下のように考えられる．
- 誤認識された記号の訓練が不十分
- OCRopus が 300dpi の画像に最適化されているため，実験で用いた 600dpi では不適当
- 後処理（言語モデリング）が正しく動作していないか未実装

④ 性能の制限
現状では OCRopus のサポートする言語が英語に限られているため，英語のみを対象としており，また読取に対応したレイアウトは 1 段組を想定している．今後は 2 段組の文書をはじめとし幅広い文書の種類に対応する必要がある．

(4) 今後の課題

本研究を通して，ディスプレイ数式の認識については対応ができたが，インライン数式については引き続き手法を開発していく必要がある．またディスプレイ数式についても精度改善のためには以下の課題を解決していく必要がある．

① インライン数式への対応
インライン数式はディスプレイ数式よりも検出が難しく，異なるアプローチが必要である．

② ディスプレイ数式の特徴量の考察
本研究で定義した特徴量が妥当なのか実験で検証する必要がある．

③ 識別アルゴリズムの検討
本システムで用いた SVM 以外に数式の識別率が高くなるアルゴリズムがないか調べなければならない．

④ 本システムを用いた定量的な実験と評価
本システムにおける数式の検出法と従来手法との比較が必要である．また，文書全体の認識率の算出，InftyReader との比較も必要である．

⑤ 実験用データセットの作成と公開
従来手法では各自の研究者が独自のデータセットを用いて実験をしているため単純に性能の比較ができない．そのため公開されて自由に使えるデータセットを用いて実験をすることが望ましい．

⑥ 数式認識システムの精度向上
行列式や表は未対応であるため，対応するために実装が必要となる．

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）

〔学会発表〕（計 3 件）

① S. Yamazaki, F. Furukori, Q. Zhao, K. Shirai, M. Okamoto : Embedding a mathematical OCR module into OCRopus, in Proc of IAPR Inter. Conf. on Document Analysis and Recognition, 19th Sep. 2011（査読あり）

② 山崎　慎平，古郡　史啓，趙　勤政，白井　啓一郎，岡本　正行：OCR ソフト OCRopus への数式認識モジュールの組込みの検討，信学技報 PRMU2010-267，Vol. 110，pp. 177-182，2011 年 3 月 11 日（査読なし）

③ 王　済凱，馬　国　峰，白井　啓一郎，岡本　正行：合焦位置を起点とした文字列抽出の一検討，信学技報 PRMU2010-286，Vol. 110，pp. 287-292，2011 年 3 月 11 日（査読なし）

〔その他〕（計 2 件）

① OCRopus 用数式認識モジュール公開ページ
http://syorserv.cs.shinshu-u.ac.jp/src/ocr/

② 数式認識ライブラリ公開ページ
http://syorserv.cs.shinshu-u.ac.jp/src/ocr/exp.html

６． 研究組織

(1)研究代表者
岡本正行（OKAMOTO MASAYUKI）
信州大学・工学部・教授
研究者番号：50109196

(2)研究分担者
白井啓一郎（SHIRAI KEIICHIRO）
信州大学・工学部・助教
研究者番号：00447723