社団法人　電子情報通信学会
THE INSTITUTE OF ELECTRONICS,
INFORMATION AND COMMUNICATION ENGINEERS

信学技報
TECHNICAL REPORT OF IEICE.

ストローク単位の確率文脈自由文法を用いたオンライン手書き数式データベースの作成

山本　遼† 　酒向　慎司† 　西本　卓也† 　嵯峨山茂樹†

† 東京大学 大学院情報理工学系研究科
〒 113-8656 東京都文京区本郷 7-3-1
E-mail: †{yamaryo,sako,nishi,sagayama}@hil.t.u-tokyo.ac.jp

あらまし　本研究では、オンライン手書き数式認識手法の性能評価とモデルの学習を目的としたオンライン手書き数式データベースの作成を行った。100 筆者による科学技術レベルの数式 200 種約 1000 データ、シンボル 245 種約 20000 データを収集した。さらにデータ内のストロークと数式内のシンボルの対応情報を自動的に推定する手法を検討し、我々の提案してきた確率文脈自由文法による数式認識手法を正解数式範囲内の文法制約下で利用することにより、半自動的なシンボルアラインメント推定を実現し、データベースのアラインメント付けの作業を大きく軽減することができた。

キーワード　数式認識 データベース 文字認識 オンライン 手書き 確率モデル 確率文脈自由文法

Making On-Line Handwritten Mathematical Expression Database Using Stroke-Based Stochastic Context-Free Grammar

Ryo YAMAMOTO†, Shinji SAKO†, Takuya NISHIMOTO†, and Shigeki SAGAYAMA†

† Graduate School of Information Science and Technology, The University of Tokyo
7-3-1, Hongo, Bunkyo-ku, Tokyo, 113-8656 Japan
E-mail: †{yamaryo,sako,nishi,sagayama}@hil.t.u-tokyo.ac.jp

Abstract　In this paper, we built an on-line handwritten mathematical expression database for evaluation and training of mathematical expression recognition systems. We collected 1000 handwriting data of 200 mathematical expressions taken from scientific papers, and 20000 handwriting data of 245 mathematical symbols. We also explored a method to automatically estimate the alignment between input strokes and symbols. We could reduce labor in tagging alignment by hand using our previously proposed method for expression recognition under the grammatical constraint of the correct expression.

Key words　Mathematical Expression Recognition, Database, Character Recognition, On-line, Handwriting, Stochastic Model, Stochastic Context-Free Grammar

1. はじめに

コンピュータへの数式情報入力は、科学技術文書の作成や数式計算ソフトウェアにおいて有用である。このための方法としては TEX, C, Matlab 等の特殊言語や、MS-Word 等に搭載されている数式エディタが利用されている。しかしこれらの方法は言語の学習や複雑な操作が必要である。一方ペンタブレットやデジタルペンによる数式入力は、普段多くの人が紙に書くのと同様に直感的で使いやすいインタフェースとなりうる。オンライン手書き数式認識問題は、それを実現するための最重要課題の一つである。

オンライン手書き数式認識問題は、入力ストローク系列の筆順、形状、2 次元構造情報を数式の構文を満たす形として認識する問題である。数多くの研究が今までなされており [1] 、多くの場合、オンライン手書き数式認識問題は、シンボル認識、2 次元構造認識、数式構文解析の 3 ステップからなる問題として捉えられている。類似した問題であるオフライン印刷数式認識においても、この枠組みは共通である。

しかしこのようなボトムアップな手法は、シンボル認識ステップにおける誤りをその後のステップで訂正できない問題がある。近年それを補う手法が盛んに提案されており、構文解析による誤り訂正手法 [2] や、数式内の文字の共起確率を考慮し

た文字認識を行う手法 [3] や、ファジー論理によるシンボルと構造の評価 [5] 等が提案されている。

これに対し我々は、シンボル認識・構造認識・構文解析を統合したストローク単位の確率文脈自由文法による認識手法 [4] を提案してきた。この手法は数式文法範囲内でシンボルと構造を同時最適化するトップダウンな手法であり、文脈情報を利用した頑健なシンボル認識が期待される。我々は [4] にて少量のデータによりその基本的な動作を確認し、より多くのデータによる性能評価を現在目指している。

しかし現在、公に入手可能な手書き数式のデータベースはほとんど存在しない。したがって認識手法の客観的・定量的な性能評価や認識手法同士の性能比較が難しく、また手書き数式スタイルの揺らぎなど実データの性質を統計的に学習することも難しい。そこで我々は、オンライン手書き数式認識の性能評価と学習を目的としてオンライン手書き数式データベースの作成を行った。本稿はこのデータベース作成における数式、デバイスの決定、また我々の提案する数式認識手法によるデータのシンボルアラインメントの自動推定手法、またそれに伴う手法の改良につい報告する。さらに得られたデータベースから、我々の提案する認識手法が数式の書き順に対して置いている仮定の妥当性について検証する。

以下 2. にてストローク単位の確率文脈自由文法による数式認識手法とその改良について、 3. にてデータベースの作成と得られたデータの分析、データのシンボルアラインメント自動推定手法について述べる。

2. ストローク単位の確率文脈自由文法を用いたオンライン手書き数式認識

2.1 手書き数式の文脈自由文法によるモデル化

オンライン手書き数式認識問題は、入力ストローク系列の形状・2 次元構造・書き順構造を数式文法を満たすパターンとして認識する問題である。 1. で述べたように、我々はこの問題をボトムアップに、すなわちシンボル認識・構造認識・数式構文解析の独立したステップにより認識するのではなく、トップダウンに、数式文法・シンボル形状・構造を考慮した手書き数式の確率モデルを用い、最尤推定により数式を認識する手法を提案している。この手法は文脈を考慮したシンボル認識を行うため、頑健性の向上が期待できる。本項ではその概要と今回新たに検討した改良点について述べる。

数式文法から生成可能な数式仮説集合を Z とし、入力手書き数式であるストローク系列を X とすると、認識問題は、$Y_0 = \arg\max_{Y \in Z} P(X|Y)$ を求める事後確率最大化問題として捉えることができる。

我々の用いる手書き数式の確率モデルは、 図 1 に示されるような文脈自由文法 (CFG: Contxt-Free Grammar) の形で表現される。この文法は一般の CFG と以下の点で異なる。

- 終端記号が入力手書き数式のストロークである。
- 非終端記号は＜項＞＜演算子＞ “x” などの数式の抽象要素であり、非終端記号 A はその位置を表すパラメタ pos_A を持つ。

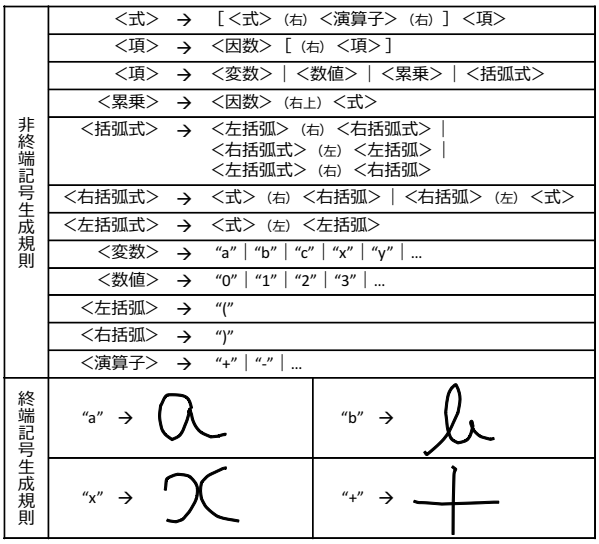

図 1　手書き数式文法の例

- 終端記号の生成規則は、$R_t = \langle A \to ab\ldots \rangle$ の形式で書かれる。 A は “x” などのシンボルを表す数式要素、$a, b, \ldots$ はストロークである。
- 非終端記号の生成規則は、チョムスキー標準形に展開することで $R_n = \langle A \to B\ \mathrm{op}_{R_n} C \rangle$ の形式で書かれる。 A, B, C はそれぞれ数式要素、op_{R_n} は (右)(右上) 等の配置演算子である。

[4] において我々は、シンボルの生成に関して “x” $\to x_{\mathrm{left}}$(右)x_{right} のように 1 ストローク単位に文法記述を行った。しかしシンボルの書き方は筆者により様々であるため、今回多筆者の手書き数式を扱うにあたり、終端記号生成規則の右辺を複数のストロークであるとした。これにより様々なシンボルの書き方を必ずしも生成規則として別々にモデル化することなく、一般的な文字認識のモデルを適用することができる。例えば＜項＞ → ＜因数＞ (右) ＜項＞という非終端記号生成規則は、「＜項＞が＜因数＞＜項＞を、(1) まず＜因数＞次いで＜項＞という書き順で、(2) ＜因数＞の右に＜項＞という論理的配置で、生成する」という意味であり、手書き数式要素間の書き順と位置関係を規定する。

2.2 確率的な生成規則の適用

生成規則は確率的に適用される。非終端記号生成規則 R_n が適用される場合、その適用確率 P_{R_n} は

$$P_{R_n} = P^{\mathrm{rew}}_{R_n} \cdot P^{\mathrm{pos}}_{R_n}$$

により表される。$P^{\mathrm{rew}}_{R_n}$ は規則の書き換え確率、 $P^{\mathrm{pos}}_{R_n}$ は配置確率である。書き換え確率は、一般的な確率文脈自由文法におけるものと同様、規則の適用される頻度を表し、数式認識問題においては数式の事前確率を与える。配置確率は数式要素の確率的な配置をモデル化するもので、規則 $R_n = \langle A \to B \mathrm{op}_{R_n} C \rangle$ の配置確率は、配置演算子 op_{R_n} に従う数式要素 A からの B, C の確率的配置

$$P^{\mathrm{pos}}_{R_n} = P(\mathrm{pos}_B, \mathrm{pos}_C | \mathrm{pos}_A, \mathrm{op}_{R_n})$$

である。

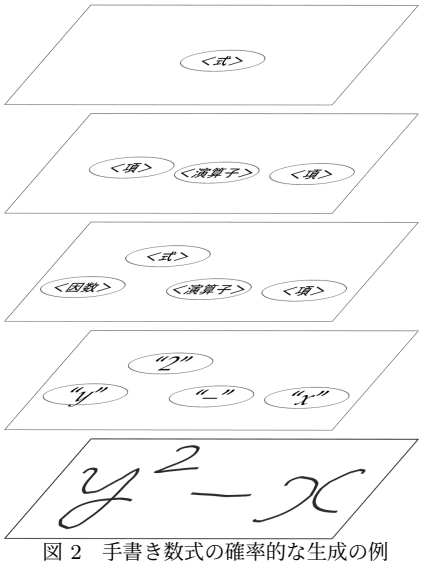

図 2　手書き数式の確率的な生成の例

終端記号生成規則 $R_t = \langle A \to ab\ldots \rangle$ の適用確率 P_{R_t} は

$$P_{R_t} = P_{R_t}^{\text{rew}} \cdot P_{R_t}^{\text{shape}}$$

のように表される。$P_{R_t}^{\text{rew}}$ は書き換え確率、$P_{R_t}^{\text{shape}}$ は

$$P_{R_t}^{\text{shape}} = P(a, b, \ldots | A)$$

で表されるストローク形状確率であり、ストロークの確率的な生成をモデル化する。

ストローク系列 X が導出

$$Y = \{R_{n_1}, R_{n_2}, \ldots, R_{n_N}, R_{t_1}, R_{t_2}, \ldots, R_{t_T}\}$$

により生成される尤度は、

$$P(X|Y) = \prod_{i=1}^{N} P_{R_{n_i}}^{\text{rew}} \prod_{i=1}^{T} P_{R_{t_i}}^{\text{rew}} \prod_{i=1}^{N} P_{R_{n_i}}^{\text{pos}} \prod_{i=1}^{T} P_{R_{t_i}}^{\text{shape}}$$

となり、各生成規則の書き換え確率・配置確率・ストローク形状確率の積となる。

2.3　生成例と認識例

この確率文脈自由文法による手書き数式の確率的生成の例を表したものが図 2 である。はじめに開始記号＜式＞がある位置パラメタをもって配置される (1 段目)。非終端記号生成規則により＜式＞は＜項＞＜演算子＞＜項＞を生成する。この規則が適用される確率は書き換え確率により与えられる。生成された＜項＞＜演算子＞＜項＞はそれぞれ特定の位置パラメタを持つが、この位置は配置確率 (この場合は (右) の配置) により与えられる (2 段目)。規則が順次確率的に適用される (3, 4 段目)。終端記号生成規則によりストローク列が生成される。この確率は書き換え確率とストローク形状確率により与えられる (5 段目)。

この手書き数式認識の枠組みは、(1) 数式生成規則、(2) 規則の書き換え (事前) 確率、(3) 配置確率モデル、(4) ストローク形状確率モデル、(5) 事後確率最大仮説の探索アルゴリズム、の設計を必要とする。[4] において我々は、(1) 括弧・積分・演算子等の概念を含む程度の複雑さの文法、(2) 仮説によらず一定の事前確率、(3) 隠れ筆記領域による配置モデル、(4)HMM によるストローク形状モデル、(5)CYK アルゴリズムによる探索を用いた。 CYK アルゴリズムによる解探索は、図 3 に示すように、はじめに部分ストローク系列に対し各シンボルの尤度を計算する。今回、終端記号生成規則の右辺を複数ストロークとしたため、第 i ストロークから第 j ストロークまでを一つのシンボルとしてストローク尤度を計算し、CYK マトリクスの (i, j) 成分に候補と尤度を記入する。以下通常の CYK アルゴリズムと同様、マトリクスの $(i, i+k)$ 成分と $(i+k+1, j)$ 成分 $(k = 0, \ldots, i-j-1)$ に記入された候補の組合せそれぞれに対し、適用可能な規則を適用した結果を (i, j) 成分に記入する。最終的にマトリクスの頂点に書き込まれた候補のうち、開始記号＜式＞であり尤度の最も高い候補が認識結果となる。以上が我々の提案する手書き数式認識手法の概要である。

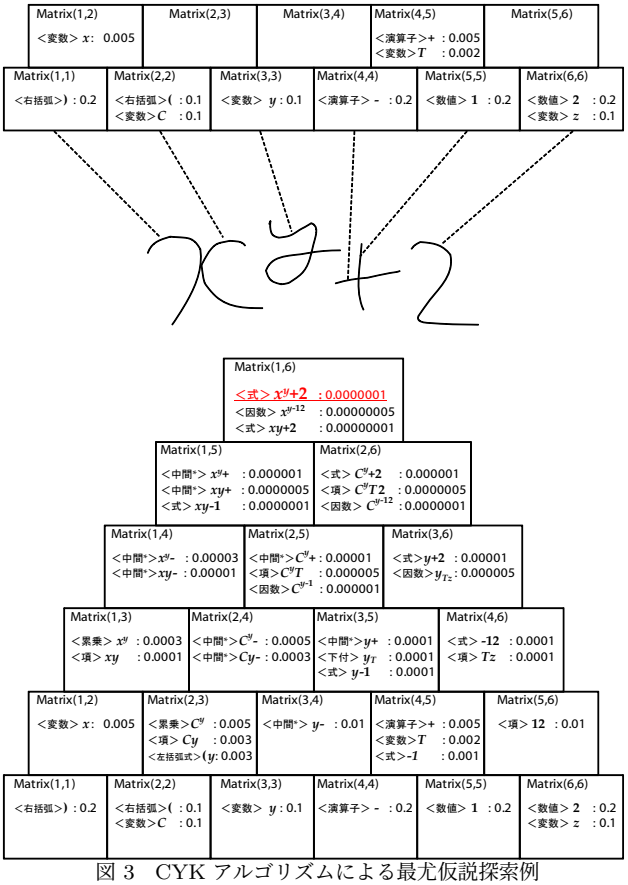

図 3　CYK アルゴリズムによる最尤仮説探索例

3.　オンライン手書き数式データベース

3.1　データベースの設計

我々は今回、提案する手書き数式認識手法の性能評価を目的として、オンライン手書き数式データベースを作成した。現在公に入手可能な手書き数式のデータベースはほとんど存在しないが、手法同士の性能比較や統計学習による頑健性の向上のため今後必要になると考えられる。

データベースを作成するにあたり、数式のドメインを決定する必要がある。現在提案されている多くの手書き数式認識の研究において、その数式ドメインはそれぞれ異なっていて、共通

表 1　各種ペン型デバイスの長所と短所

	ペンタブレット	ペンタブレット＋専用インクペン	デジタルペン	液晶タブレット
長所	安価、リアルタイム修正	安価、リアルタイム修正、書き心地が自然	安価、持ち運びが容易、書き心地が自然	リアルタイム修正、手元の文字と入力データの一致
短所	手元を見ながら書くことができない	手元の文字と入力データの不一致	手元の文字と入力データの不一致、修正が困難	高価、持ち運び困難

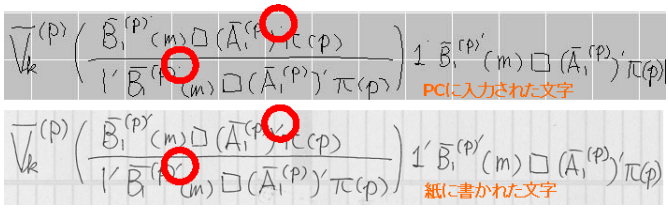

図 4　タブレットとインクペンを用いて数式を入力した時の紙上データとデジタルデータの不一致。○の部分は紙上では「ダッシュ」記号のつもりで書いたが、デジタルデータでは一点になってしまっている。このように特に「撥ね」を書いた場合や速いスピードで書いた場合にデータの不一致が起こる。

のドメインというべきものは存在しない。数式のドメインは中学・高校レベルのものから実際に論文に書かれる科学技術レベルのものまで様々な複雑さのレベルが存在する。ここで手書き数式認識技術の応用可能性として、グラフを描くなどの教育用ソフトウェア、科学技術文書や数値計算ソフトウェアへの数式入力、将来的には数式検索システムのインタフェース等があげられるが、これらの多くは科学技術レベルの数式を対象とした物である。この理由から、我々はこの度作成するデータベースも技術文書・論文レベルに準じた。また論文で実際に用いられている数式はその複雑さも当然幅広いが、複雑な数式ほどキーボードでの入力が煩雑でありペンによる入力の期待が大きいことから、我々はこれらの論文レベルの数式のうちである程度複雑なものを対象とすることとした。

以上より対象数式を、(1) 理工学系の研究論文で実際に用いられるもので、(2) ある程度複雑なものを、(3) 広い分野から集める、という方針のもと以下の手順で選択した。

（1） IEEE Transactions の全 70 論文誌から、数式を含む 200 論文をランダムに選ぶ。IEEE Transactions を論文誌として選んだ理由は、Transactions が理工学系の幅広い分野をカバーしている点、また論文のデジタルデータが入手可能な点などである。

（2） 選んだ論文それぞれから、主観で最も「複雑そうに見える」数式を抽出する。ただし極端に複雑にならないよう、長くて複数行に渡る数式やページの半分を占める巨大な行列などあまりにも複雑だと感じた場合は、その数式以外から選ぶ。

以上のように計 200 数式を得た。また同時に孤立シンボルの手書きデータ収集も行い、その対象とするシンボルは対象数式をカバーするシンボルドメインを、[2] や TEX の解説書を参考にして拡張し、全 245 シンボルとした。

表 2　収集したデータベースにおける数式書き順のバリエーション

数式構造	書き順	出現個数
(X)	(X) :	2570
(X)	( ) X :	3
(X)	X ( ) :	7
(X)	X) (:	0
$\frac{Y}{X}$	X − Y :	663
$\frac{Y}{X}$	− X Y :	490
$\frac{Y}{X}$	− Y X :	144
$\frac{Y}{X}$	Y − X :	58
$\sqrt{X}$	$\sqrt{}$ X :	154
$\sqrt{X}$	X $\sqrt{}$:	6
$\vec{X}$	$\vec{}$ X :	393
$\vec{X}$	X $\vec{}$:	17

次にデータを収集するデバイスを決定した。我々は (1) ペンタブレット、(2) タブレット用ボールペン、(3) デジタルペン、(4) 液晶ディスプレイについて検討した。結果、複雑な数式を書く場合にはペンタブレットのようにディスプレイを見ながら手元で書く方法は筆者にストレスを与える、タブレット用ボールペンなどしたい等の理由から、(4) 液晶ディスプレイを選択した。各デバイスの得失を 表 1 に示した。

3.2　データ収集と分析

筆者として 100 人の理工系の学生に入力を依頼した。入力の際の注意として、ミスをしたらアンドゥ機能でその部分を書き直してもらう事、また自然な書き順、例えば分数の分母を書いている途中で分子を書くなどしないことを被験者に求めた。この書き順の制約は、我々の提案手法において仮定されている書き順の構造である。この書き順の仮定が妥当であること、すなわち簡単なインストラクションでこの制約が守られる事を確認することも本データベース収集の目的である。ただし例外として、長い分数線と長い根号については、はじめに書いた線に「付け足して」書くことを許可した。これは入力の容易さへの配慮である。以上のように、100 筆者により数式 200 種約 1000 データ、シンボル 245 種約 20000 データを収集した。

得られたデータの書き順に関し分析を行った。得られたデータにおいて、数式の書き順は括弧、分数、平方根、アクセントをのぞいて多様性はなく、数式の先頭から順に書かれていた。これらの数式要素を書くときの書き順の統計を、1000 数式データのうち行列・ベクトルを含まない 883 数式についてとったものが表 2 である。行列・ベクトルについては対角行列や疎行列など、書き順の分析が難しいことから今回は対象外とした。

数式内の書き順に多様性のある要素について、書き順の構造が文脈自由文法で表されることが確認できる。また特に分数を書く際には、同一筆者であってもその都度違った書き順をすることがあることも観測された。数式の書き順を一意に固定せず文脈自由文法により自由度を持ったモデル化をすることは、筆者の自然な入力が実現できる点でメリットであると考えられる。

3.3　自動アラインメント推定実験

さらにデータ内のストロークと数式内のシンボルの対応情報を得ることも必要である。このシンボルアラインメントの推定

表 3　数式 $(\frac{1}{2})$ から自動生成された数式文法。正解が $(\frac{1}{2})$ であるという前提で構文解析することができる。

$(\frac{1}{2})$	→	((右) $\frac{1}{2}$)
$(\frac{1}{2})$	→	$\frac{1}{2}$) (右) (
$(\frac{1}{2})$	→	($\frac{1}{2}$ (右))
$\frac{1}{2}$)	→	$\frac{1}{2}$ (右))
$\frac{1}{2}$)	→	) (左) $\frac{1}{2}$
($\frac{1}{2}$	→	$\frac{1}{2}$ (左) (
$\frac{1}{2}$	→	1 (下) $\frac{}{2}$
$\frac{1}{2}$	→	$\frac{}{2}$ (上) 1
$\frac{1}{2}$	→	$\frac{1}{}$ (下) 2
$\frac{}{2}$	→	− (下) 2
$\frac{}{2}$	→	2 (上) −
$\frac{1}{}$	→	− (上) 1

を全て手作業で行うのは非常に大変なので、ある程度自動的にこの推定を行う事が求められる。このシンボルアラインメント推定問題は、既知である全体の数式情報を利用したシンボル認識問題である。ここで、全体の数式情報が既知であっても、シンボルの書かれる順番は表 2 のように多様性があるため、単純なストローク系列とシンボル系列のマッチング問題としては解くことができない。

我々は、このアラインメント推定問題が「正解が既知である」数式認識問題と見なせることを利用し、前項に述べた数式認識手法を応用することで解決を試みた。認識に用いる数式文法規則を「正解数式を構成する範囲」に限定することで、その数式として入力パターンを認識することができる。例えば数式 $(\frac{1}{2})$ を生成する数式文法は、表 3 に示される生成規則によって記述できる。数式データごとにこのような正解数式の範囲の文法を自動的に生成し、 2. に述べた数式認識手法によりデータのシンボルアラインメントを推定した。ここでは (1) 数式生成規則を正解数式の範囲のものに、(2) 書き換え確率はフラットとし、(3) シンボルの 2 次元的配置については考慮せず (配置確率一定とし)、(4) ストローク形状は HMM により評価し、(5) 探索は CYK アルゴリズムを用いた。ストローク形状の特徴量は、ストロークの大き縦方向に 15 ピクセルとなるように正規化し、それを横方向の 15 次元ベクトル系列とみる画像特徴量を使用し、学習データとしては同時に収集した孤立シンボルデータベースを用いた。この手法はシンボルの配置についての評価を行わないため、正解数式から考えられる全てのシンボル系列と入力ストローク系列の最適なマッチングの探索を行っていることになる。

この自動アラインメント推定の結果を図 5 に示す。我々はこの結果をベースにアラインメントの修正を行うことで、上記の 883 数式について正解のアラインメント付けを簡単に行うことができた。今後はこのデータベースによる手法の評価、データベースの公開を行う予定である。

4.　おわりに

本研究では、オンライン手書き数式認識手法の性能評価とモデルの学習を目的としたオンライン手書き数式データベースの作成を行った。100 筆者による科学技術レベルの数式 200 種約 1000 データ、シンボル 245 種約 20000 データを収集した。さらにデータベースを学習に利用するために必要であるデータ内のストロークと数式内のシンボルの対応情報を自動的に推定する手法を検討し、我々の提案してきた確率文脈自由文法による数式認識手法を正解数式範囲内の文法制約下で利用することにより、半自動的なシンボルアラインメント推定を実現し、データベースのアラインメント付けの作業を大きく軽減することができた。

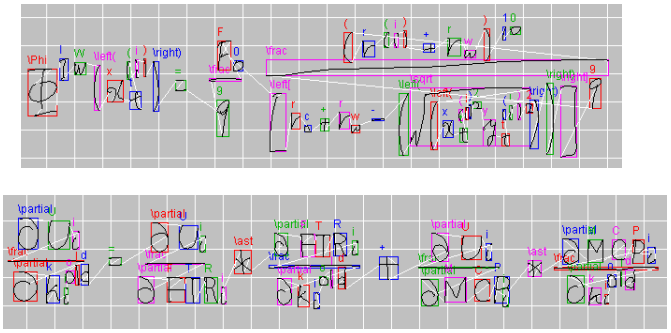

図 5　自動アラインメント推定結果例

謝　辞

本研究の一部は、東京大学大学院情報理工学系研究科戦略ソフトウェア創造人材養成プログラムの補助を受けて行われた。

文　献

[1] K. -F. Chan and D. -Y. Yeung, Mathematical Expression Recognition: A Survey , *Int. J. Document Anal. Recognit.*, vol.3, pp.3-15, Aug. 2000.

[2] U. Garain and B. B. Chaudhuri, Recognition of Online Handwritten Mathematical Expressions, *IEEE Trans. Sys. Man Cybern. Part B:Cybern.* , vol.34, pp.2366-2376, Dec. 2004.

[3] 瀧口祐介, 岡田稔, 三宅康二, 高次情報を考慮した数式文字認識の誤り訂正法の検討', 電子情報通信学会技術報告, PRMU2005-248, pp.107-112, Mar. 2006.

[4] R. Yamamoto, S. Sako, T. Nishimoto and S. Sagayama, On-Line Recognition of Handwritten Mathematical Expressions Based on Stroke-Based Stochastic Context-Free Grammar, *Proc. Int. Work. Fronteers in Handwriting Recognition (IWFHR)*, Oct. 2006.

[5] J. A. Fitzgerald and F. Geiselbrechtinger and T. Kechadi, Structural Analysis of Handwritten Mathematical Expressions Through Fuzzy Parsing, *Proc. IASTED Int. Conf. Advances in Computer Science andTechnology (ACST)*, pp.151-156, 2006