LI-009 共起確率行列を用いた数式文字認識の誤り訂正法の評価
An Evaluation of Character Recognition Error Correction Method for Mathematical Formulae using Co-occurrence Matrices

瀧口祐介 † 岡田 稔 † 三宅康二 ‡
Yusuke TAKIGUCHI Minoru OKADA Yasuji MIYAKE

1. はじめに

紙媒体に印刷された書籍や論文などの紙文書のデジタルアーカイブ化が盛んである．紙文書をデジタルアーカイブ化する場合，一般にイメージスキャナが用いられ，画像としてコンピュータに取り込まれる．取り込まれた画像はそのままでも閲覧できるが，Optical Character Recognizer (OCR) を用いれば画像中に書かれている文字を読み取ることが可能となり，高精度清書や再編集といった閲覧以外の二次利用 (知的二次利用) が可能となる．前述した事項は，数式を数多く含む科学技術系の文書の場合も同様である．しかし，一般に数式は二次元的な構造を持つため，一次元的に並んだ文字を認識対象とする OCR を用いて，これらに含まれる文字と構造を正確に認識することは困難である．この様な背景から，印刷数式を対象とした構造認識手法が報告されている．岡本らは，基本構造処理と個別構造処理を組み合わせて印刷数式の構造認識を行う手法 [1] を提案しているが，彼らは数式の構造認識における文字の誤認識に関する問題については言及していない．一方，鈴木らは文字認識の結果得られた正読文字の候補とそれらの配置を利用して正しい文字と構造を同時に決定する手法 [3] を提案している．しかし，その手法では構造を決定する為に必要な文字の誤認識には対応できないことが指摘されている．

本論文では文献 [4] で報告した文字誤認識の訂正手法を提案すると共に，本手法において正読文字の候補の数を変化させた場合の訂正結果の変動について報告し，今後の課題を述べる．提案手法は筆者らの研究室で開発を進めているオフライン数式認識理解システム [5, 6](以下，本システム) における，文字認識誤りの訂正を目的としている．本手法は訂正対象の文字を制限しないという点において，従来手法よりも優れているといえる．

2. システム構成

2.1 システム概要

本研究では，入力数式の構造を表す構造ツリーと関数名や演算の優先順位などの数学的な意味を表す意味ツリーを特に高次符号と呼び，これらを生成することをそれぞれ構造認識と意味理解と呼んでいる．本システム (図 1) は 5 つの処理部，1. 画像処理部，2. 文字認識部，3. 構造認識部，4. 意味理解部，5. 出力変換部 (それぞれの数字と図 1 中の番号は対応) と 3 つの辞書，特徴量，文字，関数辞書，によって構成されている．提案法では文字認識部で得られた文字・数学記号の候補と，構造認識部で文字認識結果とは独立に得られた数式の構造情報に注目する．更に数式構造中で隣接した文字・記号の接続方向を考慮した共起確率を利用して，文字認識結果の数式らしさを表現し，誤認識の訂正を行う．なお本論文では，数式中に現れる文字・数学記号を単に文字と呼び，文字認識部で得られる正読文字の候補を候補文字と呼ぶ．

†早稲田大学大学院情報生産システム研究科情報アーキテクチャ分野
‡中部大学工学部情報工学科

3. 文字認識部

文字認識部では，画像処理部 (図 1 の 1.) で切り出された文字領域に含まれる文字を個別に認識する．数式は，アルファベット，数学記号，ギリシャ文字および数字で構成され，それらの字体 (Typeface) と二次元的な配置で数式としての意味を表す．このため本システムにおける意味理解の精度を向上させるために，本処理部では特に字種と字体を併せて認識する．現在，文字認識部で学習している字種は，アルファベット (大文字と小文字の Roman 体及び *Italic* 体)$26 \times 2 \times 2$ 字種，数字 10 字種，数学記号 53 字種，ギリシャ文字 41 字種の全 208 字種である．文字認識には，三宅らによって開発された加重方向指数ヒストグラム [7] ($7 \times 7 \times 8 = 392$ 次元) を特徴量として用い，特徴空間における入力パタンと学習パタンの市街区距離によってクラス判別を行う．文字認識では

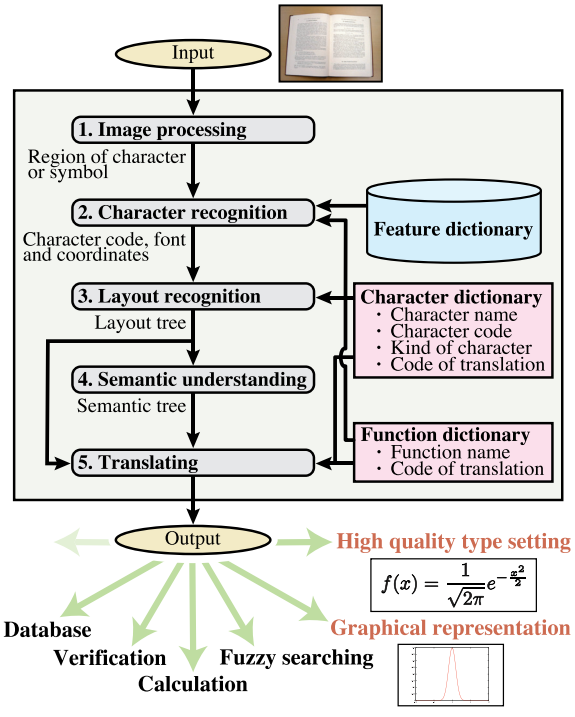

図 1: 本システムの概要
Fig. 1: Outline of the system.

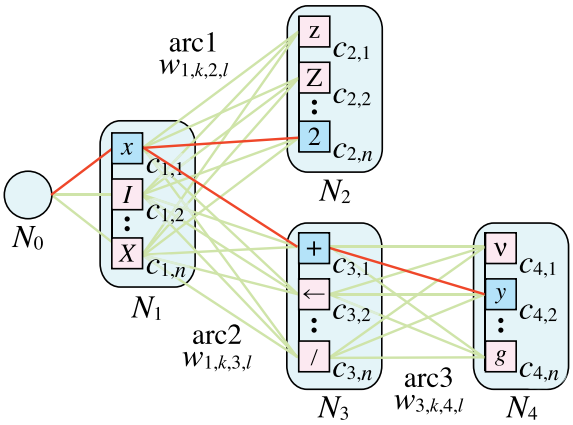

図 2: コストツリーの例 ('$x^2 + y$')
Fig. 2: A cost tree for '$x^2 + y$'.

文字領域一つに対して類似度が上位 n 個の候補文字の，文字コード，識別順位，相違度 (パタン間の市街区距離値)，書体，字体，文字領域 (候補文字の外接長方形) を得る．本処理部では入力された全ての文字領域に対して上述した認識結果を得て，構造認識部に出力する．

4. コストツリーと共起確率行列を用いた文字の誤認識の訂正

4.1 コストツリー

本提案法は，本システムの構造認識部の認識結果を利用する．まず入力画像内の各文字に対応する候補文字 n 個と構造認識部で得られた構造ツリーに基づき，コストツリーを生成する．コストツリーとは数式における構造と候補文字同士の繋がりを示したものである．ツリーのノードは入力画像内の文字に対応しており，その内部には n 個の候補文字を持つ．特に異なるノードにおける候補文字間の枝に重みを付け，その候補文字同士の繋がりの強さを表現する．図 2 にコストツリーの例を示す．ここで，図 2 の N_i はツリーの i 番目のノードを，$c_{i,k}$ は i 番目のノードに属する識別順位 k 位の候補文字を，$w_{i,k,j,l}$ は候補文字 $c_{i,k}$ と $c_{j,l}$ の間の結合の重みを表す．ただし，N_0 はツリーの仮想的なルートノードである．次に，これらを利用して候補文字 $c_{i,k}$ と $c_{j,l}$ の間のコストを与える関数 $C(i,k,j,l)$ を式 (1) の様に定義する．

$$C(i,k,j,l) = w_{i,k,j,l} D(c_{i,k}, c_{j,l}) \quad (1)$$

$$D(c_{i,k}, c_{j,l}) = dis(c_{i,k}) + dis(c_{j,l}) \quad (2)$$

$$w_{i,k,j,l} = 1 - P_{dir}(c_{i,k}, c_{j,l}) \quad (3)$$

ここで，$D(c_{i,k}, c_{j,l})$ は相違度を与える関数である．また式 2 の $dis(*)$ は文字認識部で得られた候補文字*の相違度を，式 3 の $P_{dir}(c_{i,k}, c_{j,l})$ は候補文字 $c_{i,k}$ と $c_{j,l}$ が同時に出現する確率である共起確率を示す．なお共起確率については 4.3 節で詳しく述べる．ここで N_0 をルートとして，全てのノードを一度のみ通るツリーのコストの総和が最小となるもの，つまり，

表 1: 本システムで考慮している数学キーワード
Tab. 1: Mathematical keywords considered in our system.

acos	arccos	arcsin	arctan	arg	asin
atan	constant	cos	cosec	cosh	cot
coth	csc	curl	deg	det	div
lim	exp	gcd	grad	hom	inf
ker	lg	lim	liminf	limsup	ln
log	max	min	rot	sec	sin
sinh	sup	tan	tanh		

$$\arg\min_{path} \sum_{arc \in path} C(i_{arc}, k_{arc}, j_{arc}, l_{arc}) \quad (4)$$

が，正しい文字認識の結果であることが期待される．ただし $path$ と $arc \in path$ は，それぞれ全てのノードを一度のみ通るツリーの枝の集合とその集合に含まれる枝を表し，$\arg\min_{path}$ は値を最小にする $path$ を示す．この基本概念による誤り訂正処理の流れは次の通りである．

1. コストツリーの水平方向の接続に注目して数式キーワード (4.2 節) を探索し，発見した場合はそのキーワードを構成する候補文字の相違度を減小させる．
2. 接続方向を考慮した共起確率 (4.3 節) を利用してコストツリーの枝の重み $w_{i,k,j,l}$ を決定する．
3. コストツリー内の全てのノードを一度だけ通るツリーの中から，コストの総和が最小のものを得る．

4.2 数式におけるキーワード

数式には 'sin', 'cos', 'tan', 'lim', 'log' 等の関数名や操作名などのキーワードが多く存在する．本研究では特にこれらを数式キーワードと呼び，文字の誤認識の訂正を行う際の重要な情報のひとつとして扱う．表 1 に現在のシステムで考慮している数式キーワードの一覧を示す．我々はこれらの数式キーワードを利用して，'*S*in', 'C0*S*', '1im', '1Og' 等の様な入力された数式画像に含まれる数式キーワード中に生じる文字の誤認識の訂正を行う．具体的には，コストツリー内で水平方向に接続されたノードに注目し，それらに含まれる候補文字の組合せによってキーワードの構成が可能かを調べる．そしてキーワードの構成が可能であれば，キーワードを構成する全ての候補文字の相違度を軽減させる．

4.3 文字の接続方向別の共起確率行列

コストツリーにおける重み $w_{i,k,j,l}$ を定義するために，隣接文字の接続方向別の共起確率行列を利用する．図 3 で示す様に，提案法では数式中の文字の接続方向として，7 方向を仮定している．なお本システムでは $\sum_{i=1}$ の様な，注目文字に複数の文字で構成された式が隣接する場合，その式の先頭文字に接続する．この例では $\sum$ の下方向に i が繋がる．共起確率行列 $\boldsymbol{P}_{dir} = \{P_{dir}(c_{i,k}, c_{j,l})\}$ は，

$$P_{dir}(c_{i,k}, c_{j,l}) = \begin{cases} \dfrac{\mathrm{num}_{\mathrm{to}}(dir, c_{i,k}, c_{j,l})}{\mathrm{num}_{\mathrm{from}}(dir, c_{i,k})} & \mathrm{num}_{\mathrm{from}}() > 0 \\ 0 & \mathrm{num}_{\mathrm{from}}() = 0 \end{cases}$$

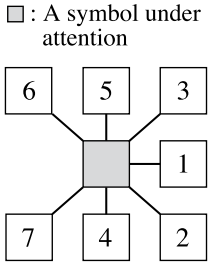

図 3: 文字の接続方向
Fig. 3: Directed connections of adjacent symbols.

表 2: InftyCDB-1[8] における接続方向別の出現度数
Tab. 2: Total numbers of the directed connections of InftyCDB-1[8].

番号	接続方向	出現度数
1	水平	101,133
2	右下	14,446
3	右上	8,440
4	下	4,084
5	上	1,390
6	左上	10
7	左下	2
合計		129,505

表 5: 訂正結果の分類
Tab. 5: Classification of correction results.

(a) 結果の分類

訂正結果	内容	集合
成功	正読 → 正読	S_s
	誤読 → 正読	S_f
失敗	誤読 → 誤読	F_f
	正読 → 誤読	F_s

(b) 失敗要因の分類

内容	集合
要因 α による失敗	A
要因 β による失敗	B
その他の要因による失敗	C

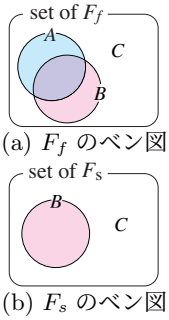

図 4: 失敗結果のベン図
Fig. 4: Venn diagrams of failure results.

で定義される共起確率を要素に持つ，非対称正方行列である．ここで $dir \in \{1, 2, ..., 7\}$ はコストツリー中の候補文字 $c_{i,k}$ から $c_{j,l}$ への接続方向を表し，$\mathrm{num}_{\mathrm{from}}()$ と $\mathrm{num}_{\mathrm{to}}()$ は，それぞれ接続元の候補文字の出現数と接続元と接続先の候補文字が同時に生起した場合の出現数を表す．このため，$0 \leq P_{dir}(c_{i,k}, c_{j,l}) \leq 1$ と $P_{dir}(c_{i,k}, c_{j,l}) \neq P_{dir}(c_{j,l}, c_{i,k})$ をそれぞれ満たす．

5. 実験結果と考察

5.1 共起確率行列の生成

鈴木らによって提供されている数式画像の正解付きデータベース InftyCDB-1 [8] を利用して，数式における 7 種類の方向別共起確率行列を生成した．InftyCDB-1 に含まれる文字の総字種数は 1,561 であるため，生成された共起確率行列は 1,561 行 1,561 列となる．本研究で調査した同データベース中の 20,097 個の数式についての，文字同士の接続方向別の出現度数を表 2 に示す．

5.2 誤り訂正

本実験は約 50 年程前に印刷された物理学の書籍 [9] から取り込んだ 59 枚の数式画像，1,330 個の文字について行った．ただし本実験では提案手法のみの評価を目的としているため，数式の構造情報は手作業によって正解を入力している．特に提案手法における候補文字数の影響を調べるため，文字認識部で出力する候補文字数を 1〜8 個まで変化させて実験を行った．実験の結果，提案手法によって文字単位の認識率が約 87.4% から 92.4% に，数式単位の認識率が約 10.0% から 26.7% に改善したことを確認した．ただし文字単位および数式単位の認識率とはそれぞれ，総文字数に対する字種と字体が正しく認識された文字数の割合と，総数式数に対する数式に含まれる全ての文字が正読である数の割合である．実験によって得られた誤り訂正前の文字認識率を表 3 に，候補文字数を変化させた訂正結果を表 4 に示す．さて，訂正の主要な失敗要因は次の通りである．

- **要因 α**: 正読文字が候補文字の中に存在しない
- **要因 β**: 共起確率の偏り

これに基づいて実験結果の成功と失敗を表 5 に示す様に分類し，分類別の集合の推移を示したグラフをそれぞれ図 5, 6 に示す．なお，表 5 による実験結果の集合の関係は図 4 の様に表される．ただし本実験では，$P_{dir}(c_{i,k}, c_{j,l}) > 0.1$ となるものを**要因 β** として分類した．

5.3 考察

図 5 より候補文字が 6 個以上の結果では認識率の変動は見られない (**事実 1**)．これは提案手法におけるコストの定義 (式 2) に候補文字の相違度を用いているためであると考えられる．同様に図 6 の total と $F_f \cap B$, $F_f \cap A \cap B$, $F_s \cap B$, $F_s \cap C$ (total 以外は図中で重なって表示) も変動は見られないが，候補文字の個数の増加に伴い $F_f \cap A$ と $F_f \cap C$ の割合は，それぞれ減少あるいは増加し続け

表 3: 文字認識結果 (誤り訂正無し，総数: 1,330 文字)
Tab. 3: Character recognition result (without error correction, total: 1,330 symbols).

識別順位	1	2	3	4	5	6	7	8	9	10	11	12	13 以下
正読数	1,162	97	26	10	10	3	7	5	5	1	0	2	2
認識率 [%]	87.4	7.3	2.0	0.8	0.8	0.2	0.5	0.4	0.4	0.1	0.0	0.2	0.2
累積認識率 [%]	87.4	94.7	96.6	97.4	98.1	98.3	98.9	99.2	99.6	99.7	99.7	99.8	100.0

表 4: 訂正結果の候補文字数における結果の推移 (総数: 1,330 文字)
Tab. 4: Experimental results for the case of change of a number of candidates (total: 1,330 symbols).

候補文字数	1	2	3	4	5	6	7	8
成功率 [%]	87.4	90.2	91.4	92.0	92.4	92.4	92.4	92.4
失敗率 [%]	12.6	9.8	8.6	8.0	7.6	7.6	7.6	7.6

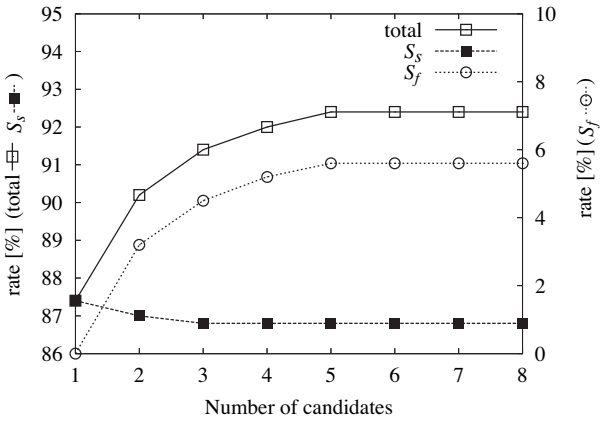

図 5: 成功結果の推移 (右軸: S_f, 左軸: その他)
Fig. 5: Changes of successful results.

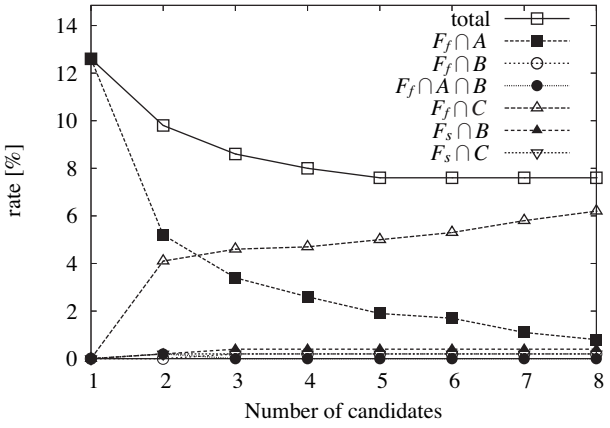

図 6: 失敗結果の推移
Fig. 6: Changes of failure results.

ている (**事実 2**). 特に $F_f \cap A$ の要素が $F_f \cap C$ に移っていることから，$F_f \cap C$ の多くは相違度の高い認識結果であると考えられる．従って本手法を用いる場合，**事実 1** より，訂正に用いる候補文字を適当な個数 (例えば 6 個) 以上にする必要はない．また**事実 2** より，文字認識精度を向上させることで候補文字個数のより少ない段階で $F_f \cap A$ の全てが $F_f \cap C$ に移ると思われる．このため，高精度な文字認識処理の実装と，それによる失敗結果の分類と対応策の検討が今後の課題として挙げられる．

6. まとめ

本論文では，数式を対象とする文字認識の誤り訂正手法について提案した．提案手法では，構造認識処理で得られた数式の構造情報に基づき，数学関数名等の数式キーワード情報と接続方向別に学習した文字の共起確率を要素として持つ共起確率行列を利用した．まず，文字認識処理部の結果として得られた候補文字と数式の構造情報からコストツリーを生成した．その後，コストツリー中の全てのノードを一度だけ通るツリーにおけるコストの総和を計算し，そのコストが最小となるものを正しい認識結果として扱った．1,330 文字を含む 59 枚の数式画像を用いた実験の結果，提案法によって文字単位の文字認識率が 87.4% から 92.4% に，数式単位の認識率が 10.0% から 26.7% に改善されたことを確認した．また提案法において候補文字数を変化させた場合の結果の変化を調査した．調査の結果，一定数以上の候補文字は必要ない事と，より高精度な文字認識処理の必要性を確認した．なお，現在のシステムにおける文字認識では認識対象 1 字種あたり 1 サンプル強しか学習していないため，学習サンプル数を増加させることで，文字認識精度の改善が見込まれる．また提案手法は数式の構造情報が誤っている場合，正しい訂正を期待できない．そのため文字認識の結果に依存しない，高精度な構造認識手法の開発が必要である．これらの改善と開発が今後の課題である．

なお，本論文で紹介した共起確率行列は n 個の累乗あるいは積により，数式中で隣接する n 文字の同時生起確率を表現できる．これを利用すれば数式中で離れた文字の共起確率を求めることも可能である．そこで，これを利用した提案手法の拡張法についても検討する．

参考文献

[1] 岡本正行, 東裕之: “記号のレイアウトに注目した数式構造認識”, 信学論 D–II, Vol. J78–D–II, No. 3, pp. 474–482 (1995)

[2] H. J. Lee and J. S. Wang: “Design of a Mathematical Expression Understanding System”, *Pattern Recognition Letters*, Vol. 18, No. 3, pp. 289–298 (1997)

[3] Y. Eto and M. Suzuki: “Mathematical Formula Recognition using Virtual Link Network”, *Proc. of ICDAR 2001 - 6th Int’l Conf. on Document Analysis and Recognition*, pp. 762–767 (2001)

[4] 瀧口祐介, 岡田　稔, 三宅康二: “高次情報を利用した数式文字認識の誤り訂正法の一検討”, 信学技報, PRMU2005-248, pp. 107–112, 6 pages (2006)

[5] Y. Chen and M. Okada: “Structural Analysis and Semantic Understanding for Offline Mathematical Expressions”, *Int’l J. of Pattern Recognition and Artificial Intelligence*, Vol. 15, No. 6, pp. 967–987 (2001)

[6] Y. Takiguchi, M. Okada and Y. Miyake: “A Fundamental Study of Output Translation from Layout Recognition and Semantic Understanding System for Mathematical Formulae”, *Proc. of ICDAR 2005 - 8th Int’l Conf. on Document Analysis and Recognition*, pp. 745–749 (2005)

[7] 鶴岡信治, 栗田昌徳, 原田智夫, 木村文隆, 三宅康二: “加重方向指数ヒストグラム法による手書き漢字・ひらがな認識”, 信学論 D, Vol. 70-D, No. 7, pp. 1390–1397 (1987).

[8] M. Suzuki, S. Uchida and A. Nomura: “A Ground-truthed Mathematical Character and Symbol Image Database”, *Proc. of ICDAR 2005 - 8th Int’l Conf. on Document Analysis and Recognition*, pp. 675–679 (2005)

[9] D. H. Menzel: “Fundamental Formulas of Physics”, Prentice-Hall, Inc., (1955)