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.

# 数学文書データベースの解析

内田　誠一†　　野村　明弘††　　鈴木　昌和†††

† 九州大学大学院システム情報科学研究院
†† 九州大学大学院数理学府
††† 九州大学大学院数理学研究院
〒 812–8581 福岡市東区箱崎 6–10–1
E-mail: †uchida@is.kyushu-u.ac.jp, ††† suzuki@math.kyushu-u.ac.jp

**あらまし**　数式を含んだ文書のための実用的な OCR の実現のために，数学文書を幾つかの観点から解析する．具体的には，67 万文字からなる正解付き数学文書データベースを準備し，(i) 文字カテゴリ数，(ii) 接触文字や分離文字などの異常文字数，(iii) 文字サイズの変動量，(iv) 数式の複雑さ，の 4 点を中心として定量的に解析する．解析を通して，数学文書を認識する際の困難性が数値として明らかにする．また，そうした問題点に対する解決策についても触れる．
**キーワード**　数学文書, OCR, データベース, 定量的解析

# Quantitative Analysis of Mathematical Documents

Seiichi UCHIDA†, Akihiro NOMURA††, and Masakazu SUZUKI†††

† Faculty of Information Science and Electrical Engineering, Kyushu University
†† Graduate School of Mathematics, Kyushu University
††† Faculty of Mathematics, Kyushu University
Hakozaki 6–10–1, Higashi-ku, Fukuoka-shi, 812–8581 Japan
E-mail: †uchida@is.kyushu-u.ac.jp, ††† suzuki@math.kyushu-u.ac.jp

**Abstract**　Mathematical documents are analyzed from several viewpoints to develop practical OCR for mathematical and other scientific documents. Specifically, the following four viewpoints are quantified using a large-scale database of mathematical documents, which contains manually ground-truthed 670,000 characters: (i) the number of character categories, (ii) abnormal characters (e.g., touching characters), (iii) character size variation, and (iv) the complexity of math expressions. The result of those analyses clarifies the difficulties on recognizing math documents and then suggests the promising directions to overcome them.
**Key words**　mathematical documents, OCR, database, quantitative analysis

## 1. ま え が き

数学や工学の分野における文書画像を対象とした OCR [1]（以下，数式 OCR）とは，それら文書の画像中の文字や記号ならびに数式の構造を解析・認識し，最終的に XML や LaTeX, MathBraille [2] (数式記述点字) といった数学記述言語を出力するシステムである．数式 OCR が実現すれば，数学文書の蓄積や検索等が容易になる．このため，科学文書の電子図書館化 [3], [4] の必須技術として検討されている．

本論文では，実用的な数式 OCR を実現するための基礎的検討として，その処理対象である数学文書を幾つかの視点から解析する．具体的には，全体で 450 ページ以上から構成される正解付きの英語数学文書データベースを作成し，それを

（1） 各カテゴリの文字数
（2） 接触文字や分離文字などの異常文字
（3） 文字サイズの変動
（4） 数式の複雑さ

の 4 点を中心として定量的に解析する．さらにこの解析を通して，数式 OCR 実現の際の問題点およびそれに対する解決案を明らかにする．

以下で使用する用語について説明する．まず「文字」は，“A” など通常の文字に加え，“+” などの数学記号も指す．また「カテゴリ」は文字種別の最小単位を指し，「タイプ」はフォントなど共通の性質を持つカテゴリの集合のことを指す．例えば，“A”，“B”，“C” は同じ Roman というタイプに属する 3 つのカテゴリであり，“A”，“$A$”，“$\mathcal{A}$”，“$\mathbb{A}$”，“$\mathfrak{A}$”，“$\mathcal{A}$” は 6 つの異な

表 1　データベース中の文字の内訳.

Table 1　Contents of database.

| type | category examples | #predefined categories | text region #cat. | text region #char ( %) | math region #cat. | math region #char ( %) | total #cat. | total #char ( %) |
|---|---|---|---|---|---|---|---|---|
| accent | ˆ ˜ ¯ ¨ ˙ ˇ _ | 13 | 1 | 2 ( <0.01) | 7 | 2,611 ( 1.73) | 7 | 2,613 ( 0.39) |
| arrow | ←↔←↖ | 16 | 2 | 6 ( <0.01) | 5 | 1,083 ( 0.72) | 5 | 1,089 ( 0.16) |
| big symbol | ∑∫∏ | 16 | 0 | 0 ( 0.00) | 9 | 1,243 ( 0.82) | 9 | 1,243 ( 0.19) |
| blackboard bold | 𝔸𝔹ℂ𝔻𝔼𝔽 | 26 | 0 | 0 ( 0.00) | 9 | 425 ( 0.28) | 9 | 425 ( 0.06) |
| calligraphic | $\mathcal{A}\,\mathcal{B}\,\mathcal{C}\,\mathcal{D}\,\mathcal{E}\,\mathcal{F}$ | 26 | 0 | 0 ( 0.00) | 19 | 591 ( 0.39) | 19 | 591 ( 0.09) |
| German | $\mathfrak{A}\,\mathfrak{B}\,\mathfrak{C}\,\mathfrak{a}\,\mathfrak{b}\,\mathfrak{c}$ | 52 | 0 | 0 ( 0.00) | 24 | 1,022 ( 0.68) | 24 | 1,022 ( 0.15) |
| Greek | ΓΔΘ$\alpha\beta\gamma$ | 39 | 4 | 24 ( 0.00) | 36 | 12,508 ( 8.27) | 36 | 12,532 ( 1.87) |
| Italic | *A B C a b c* | 61 | 56 | 62,594 ( 12.07) | 52 | 49,938 ( 33.01) | 57 | 112,532 ( 16.79) |
| extended Latin | Ä Æ è *Ä Æè* | 106 | 31 | 409 ( 0.08) | 3 | 13 ( 0.01) | 32 | 422 ( 0.06) |
| numeric | 0 1 2 *0 1 2* | 20 | 20 | 12,332 ( 2.38) | 14 | 15,152 ( 10.01) | 20 | 27,484 ( 4.10) |
| operator | + − ×/ < & | 90 | 5 | 72 ( 0.01) | 47 | 19,944 ( 13.18) | 48 | 20,016 ( 2.99) |
| others | # % ∞ ∀ ∃ † | 50 | 13 | 3,449 ( 0.67) | 20 | 4,893 ( 3.23) | 25 | 8,342 ( 1.24) |
| parenthesis | ( ) { } [ ] | 18 | 7 | 7,923 ( 1.53) | 11 | 29,722 ( 19.65) | 11 | 37,645 ( 5.62) |
| point | , . ‘ ’ | 13 | 7 | 20,631 ( 3.98) | 8 | 6,566 ( 4.34) | 11 | 27,197 ( 4.06) |
| Roman | A B C a b c | 61 | 56 | 411,311 ( 79.29) | 54 | 5,567 ( 3.68) | 56 | 416,878 ( 62.22) |
| script | $\mathcal{A}\,\mathcal{B}\,\mathcal{C}\,\mathcal{D}\,\mathcal{E}\,\mathcal{F}$ | 26 | 0 | 0 ( 0.00) | 2 | 7 ( <0.01) | 2 | 7 ( <0.01) |
| total | | 633 | 202 | 518,753 (100.00) | 320 | 151,285 (100.00) | 371 | 670,038 (100.00) |

注: (1) Roman と Italic の 2 タイプには，それぞれ “fi” のような合字が 9 カテゴリ含まれている．

(2) タイプ extended Latin は 53 種の拡張ラテン文字について，立体と斜体がそれぞれ用意された結果，106 カテゴリとなっている．

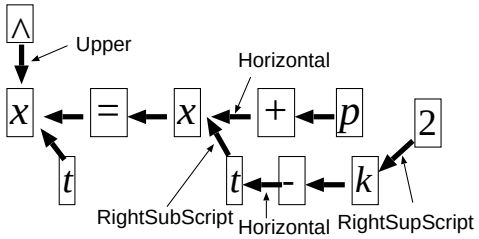


図 1　数式 “$\hat{x}_t = x_{t-k^2} + p$” の構造を表現するリンク.

Fig. 1　Links for “$\hat{x}_t = x_{t-k^2} + p$”.

るタイプ (それぞれ Roman，Italic，calligraph，blackboard bold，German，script) に属するカテゴリである．本論文で仮定するカテゴリとタイプの詳細については後述する．それぞれの文字は「テキスト領域」と「数式領域」のいずれかに属する．数式領域には，式番号がついたような数式だけでなく，行中の数式（インライン数式）も含む．例えば “The term $x^2$ equals to $y^2 + z^2$.” という表現において，“$x^2$” と “$y^2 + z^2$” は数式領域であり，それ以外はテキスト領域である．また，この例にある “$x$” や “+” を「ベースライン文字」，“2” を「添字」と呼ぶ．分数 $\frac{a^2}{b}$ においては，“$a$” と “$b$” をベースライン文字とし，“2” を添字とする．

## 2.　データベースの概要

### 2.1　データ収集

本論文で用いたデータベースに収録した文書は，純粋数学に関する 29 編の英語論文であり，発行年度は 1970∼2000 年である．文書の選出は基本的にランダムだが，ほとんど数式を含まないような文書は避けた．付録 1. に，それら論文のリストを示す．総ページ数は 453，総文字数は約 67 万であった．このデータベースは，数式 OCR に関する従来の検討において使用されているものに比べ，相当大規模であると言える．(例えば文献 [5] では約 15,000 文字，文献 [6] では約 10,000 文字からなるデータベースを使用している.）なお，行列，表，図の領域については除外した．

すべてのページ画像は，商用スキャナ (RICOH imagio Neo 450) を用いて，600dpi で撮像され二値化された．その際の二値化レベルはスキャナが自動設定したものであるが，紙や印刷の悪い文書では接触文字や分離文字などの異常文字が散見された．

### 2.2　正解付け作業

すべての文字について，カテゴリなどの属性情報（いわゆる ground truth）を手動で付与した．この正解付け作業は数学を専攻する 7 人の学生を中心として行われた．付与した主な属性情報は以下の通りである．

（1） タイプとカテゴリ
（2） 数式領域とテキスト領域の別
（3） 正常文字と異常文字の別
（4） サイズ（高さと幅）
（5） ページ中の位置
（6） 添字のレベル
（7） 隣接文字との位置関係を表すリンク

正解付け作業の際にあらかじめ定義していたタイプとカテゴリの数はそれぞれ 16 と 633 であった．これら 16 タイプと各タイプに属するカテゴリの例を表 1 に示す．数学文書以外にもよく見られるタイプに加え，big symbol など数学文書に独特なタイプも準備した．なお，この作業において太字 (bold) の区別はしなかった．従って，太字というタイプはなく，“**A**” と “A” は

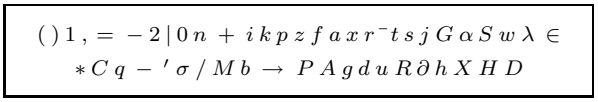

図 2　数式領域における頻出カテゴリ 1 位 (“(”) から 50 位 (“$D$”)．ここで “=” の次の “–” はマイナス，“$r$” の次の “¯” はオーバーライン，“$q$” の 次の “–” は分数線．
Fig. 2　Top 50 categories with high frequencies in math region.

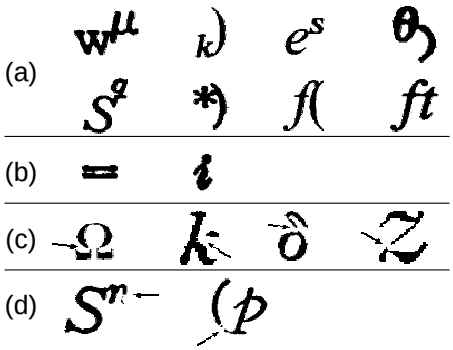


図 3　(a) 接触文字，(b) 自己接触文字，(c) 分離文字，(d) 分離を伴う接触文字．矢印は分離個所を示す．
Fig. 3　(a) Touching characters, (b) self-touching characters, (c) broken characters, and (d) touching and broken characters. Broken points are indicated by arrows.

同じカテゴリ “A”(Roman) として扱われる．これは太字とそうでない文字の違いは微妙なことが多く，またその違いは文書毎で異なるためである．太字であるか否かは数式においてしばしば重要な意味を持つので，その取り扱いについて今後改めて検討する予定である．

第 5 の属性である添字のレベルとは，数式領域の文字に付与される属性である．レベル 0 をベースライン文字とする．レベル 1 は通常の添字 (数式 “$x^{(y+z)^2}$” の “$(y+z)$”)，レベル 2 は二重添字 (同数式の “2”) を表す．

リンクとは，数式の構造を表現するために付与された属性である．図 1 に例示したように，リンクは隣接する文字の位置関係を表すものであり，このリンクにより 1 つの数式全体の構造は木として表現される．リンクの種類は，水平 (Horizontal)，右上付き (RightSupScript)，右下付き (RightSubScript)，左上付き (LeftSupScript)，左下付き (LeftSubScript)，上付き (Upper)，下付き (Under) の 6 種である．水平以外のリンクを含むような数式は，２次元構造を持つといえる．

## 3.　カテゴリ数およびタイプ毎の文字数

### 3.1　カテゴリ数

表 1 に，本データベースにおけるタイプ毎のカテゴリ数および文字数を示す．この表から，数学文書が通常の文書に比べて非常に多くのカテゴリ (371) から成り立っていることがわかる．実際，テキスト領域に限定するとカテゴリ数は 202 であり，これが通常の文書のカテゴリ数とほぼ同じと考えると，数学文書は 2 倍弱のカテゴリからなっていることがわかる．従って数学文書を現実的な時間で認識するためには，通常文書の識別処理よりも計算量の少ないものが要求されることがわかる．

表 2　異常文字数. 上段：文字数．下段：該当領域内における割合 (%).
Table 2　The number of abnormal characters.

| | | normal | touched | self-touched | broken | touch& broken | total |
|---|---|---|---|---|---|---|---|
| text | | 510,930 | 6,083 | 327 | 1,401 | 12 | 518,753 |
| | | 98.49 | 1.17 | 0.06 | 0.27 | <0.01 | 100.00 |
| math | base-line | 112,379 | 1,212 | 66 | 754 | 9 | 114,420 |
| | | 98.22 | 1.06 | 0.06 | 0.66 | 0.01 | 100.00 |
| | script | 35,595 | 291 | 254 | 725 | 0 | 36,865 |
| | | 96.55 | 0.79 | 0.69 | 1.97 | 0.00 | 100.00 |
| | sub-total | 147,974 | 1,503 | 320 | 1,479 | 9 | 151,285 |
| | | 97.81 | 0.99 | 0.21 | 0.98 | 0.01 | 100.00 |
| total (text+math) | | 658,904 | 7,586 | 647 | 2,880 | 21 | 670,038 |
| | | 98.34 | 1.13 | 0.10 | 0.43 | <0.01 | 100.00 |

表 3　異常文字含有率により全 29 文書を分類した結果.
Table 3　Distribution of abnormal character rates of all 29 documents.

| abn. rate(%) | <0.2 | 0.2∼0.5 | 0.5∼1 | 1∼2 | 2∼3 | 3∼4 | 4∼5 | >5 |
|---|---|---|---|---|---|---|---|---|
| #documents | 3 | 4 | 5 | 6 | 3 | 5 | 1 | 2 |

### 3.2　タイプ毎の文字数

イタリック文字の認識の重要性も表 1 からわかる．イタリック文字は傾いており，通常の OCR ではセグメンテーションの失敗などを理由に誤認識となりやすい．従って数式 OCR では，イタリック文字に特化した処理を準備すべきと考えられる．なお，イタリック文字は数式領域において変数や関数を表す際に多用されるが，表 1 からテキスト領域にも多数含まれていることがわかる．これは数学文書においては，定理や定義を記述した部分がイタリック体で印字されていることが多いためである．

拡張ラテン文字はデータベース中に 422 文字含まれ，主に参考文献の著者名の箇所に存在した．数式領域にも 13 個含まれているが，人物名に由来した特殊な変数名や関数名であった．

### 3.3　カテゴリの頻度

数式領域において高頻度だった上位 50 カテゴリを図 2 に示す．イタリック文字だけでなく，括弧類やアクセント類，演算子も含まれていることがわかる．

数学文書中には非常に良く似たカテゴリの文字が存在する．例えば，カテゴリ “r”，“$r$”，“$\gamma$”，“$\Upsilon$”，“$\mathfrak{r}$” は類似しており，従って互いに誤認識となりうる．こうした状況の下で，全体として認識率を向上するためには，各カテゴリの発生頻度を事前確率として識別の際に用いる方針が考えられる．このようにすることで，高頻度カテゴリの文字（上の例では “r” や “$r$”）が低頻度カテゴリに誤認識される可能性を減らすことができる．ところで，表 1 からわかるように，低頻度カテゴリの文字（上の例では “$\Upsilon$” や “$\mathfrak{r}$”）の頻度は，ほぼ 0 と言ってよいほど低い．よって，低頻度カテゴリの文字が数式を理解する上で重要なものであっても，この事前確率を用いる限りそれらが正しく認識される可能性は極めて低くなる．より見込みのある他の方針としては，各文書に依存した事後処理の導入が考えられる．例えば，ある文書においてカテゴリ “r” に認識されたものをすべて集めて相互比較するという事後処理を行う．文書内で “r” の形

表 4　数式領域における接触文字／分離文字数の上位 15 カテゴリ．各表の中段 (#char) はその異常文字の個数，下段 (#doc) はその異常文字が含まれた文書数．ここで接触文字の 11 位はオーバーライン，14 位は 分数線．

Table 4　Top 15 categories with many abnormal characters in math region.

Touching:

| cat. | ( | $p$ | ) | $\partial$ | r | 2 | $v$ | $\rho$ | $r$ | $V$ | ¯ | $M$ | i | – | $w$ |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| #char | 251 | 157 | 143 | 62 | 61 | 45 | 43 | 41 | 33 | 32 | 30 | 29 | 26 | 25 | 25 |
| #doc | 8 | 9 | 10 | 1 | 3 | 6 | 2 | 3 | 2 | 3 | 5 | 1 | 2 | 4 | 3 |

Broken:

| cat. | $s$ | $p$ | K | $\mathfrak{N}$ | $Y$ | $\prod$ | $j$ | $k$ | $n$ | ( | 2 | $\cong$ | $a$ | $x$ | ) |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| #char | 143 | 129 | 78 | 68 | 61 | 56 | 54 | 49 | 48 | 37 | 34 | 28 | 28 | 27 | 23 |
| #doc | 3 | 4 | 1 | 1 | 2 | 1 | 2 | 6 | 7 | 14 | 12 | 6 | 2 | 7 | 8 |

状は一定であるので，大多数とは形状の異なるものとして “r” に誤認識されている文字（例えば ‘ϒ”）を検出することができるものと考えられる．

## 4. 異 常 文 字

### 4.1 異常文字の種類と分布

本節では，誤認識の発生源となる異常文字について，幾つかの視点から解析する．図 3 に例示するように，異常文字は，接触文字，自己接触文字，分離文字，接触かつ分離文字，の 4 種に分類される．ここで自己接触文字とは，もともと連結成分が 2 個以上存在し，それらのうち幾つかが接触した文字である．

表 2 に示すように，データベースには 11,134 個の異常文字があり，これは全文字の 1.66%にあたる．このように，異常文字は数式 OCR を実現する上で無視できない程度含まれている．また，全文字の 1.13%が接触文字，0.10%が自己接触文字，0.43%が分離文字，0.01%未満が接触かつ分離文字であった．領域ごとで見ると，数式領域には 2.19%の確率，テキスト領域には 1.51%の確率で異常文字が存在しており，従って数式領域の方が異常文字の存在確率が高いことがわかる．

表 3 は異常文字含有率で全 29 文書を分類した結果である．このように文書によって異常文字含有率は様々であることがわかる．これは印刷環境 (紙質，印字濃度，フォント，文字間スペースなど) が各文書で異なるためと考えられる．一方，1 文書内に限定すれば，印刷環境は一定していると言える．その結果，後述するように，１文書内に特定の異常文字 (例えば “$\partial$” の接触文字) が多発することがある．以上から，異常文字の検出/正常化処理においても，文書依存処理が重要であると言える．

表 4 は，数式領域において接触文字および分離文字が多かった上位 20 カテゴリである．数式領域の認識率を向上させるためには，これらカテゴリの異常文字を確実に認識するような仕組が必要であると言える．同表にはその異常文字が含まれた文書数も示している．丸括弧 “()” の接触文字など複数の文書で散見されるものもあるが，“$\partial$” や “$M$” の接触文字，および “K”，“$\mathfrak{N}$”，“$\prod$” の分離文字については，1 つの文書のみに集中して存在するものも多い．上述したように，こうした集中は印刷環境が一定であることに因って発生する．

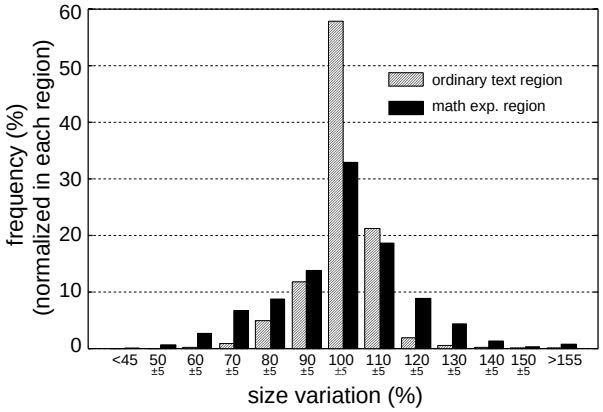


図 4　カテゴリ平均サイズからの変動量の分布．

Fig. 4　Frequency of size variation from the average of each category.

### 4.2 数式領域の接触文字の解析

最も多い異常文字である接触文字は，通常の文書 OCR では致命的な問題にはならない．実際，テキスト領域に接触文字があったとしても，セグメンテーションや単語辞書の効果により正しく認識されることが多い．しかしながら，数式領域においては，接触文字は誤認識となりやすい．これは，数式においては定まった単語辞書が存在せず，さらに図 3 の例にもあるように，水平方向以外の接触も起こりうるため，正確なセグメンテーションが困難となるためである．もし数式領域中の接触文字 (1,503 文字) がすべて誤認識となったとすると，数式領域の認識率は最高でも 99%に留まることになり，さらには数式構造の解析成功率にも大きな悪影響を及ぼすと考えられる．従って，数式領域の接触文字の検出ならびにその分離手法の開発は，高精度な数式 OCR を実現する上で非常に重要な課題であることがわかる．

数式領域においては，水平以外の方向，すなわち垂直方向と対角方向に接触した文字が多数存在した．具体的には，数式領域中の全 761 接触文字対のうち，146 対（約 20%) はベースライン文字と添字文字による接触であり，従って水平以外の方向の接触文字対であった．この結果は，こうした非水平接触文字の検出・セグメンテーション手法 [7], [8] の必要性を裏付けている．

多重接触もわずかながら存在した．数式領域には 3，4 文字の接触がそれぞれ 14，6 例あった．なおテキスト領域には 3，4，5，6，7 文字の接触がそれぞれ 155，25，11，1，5 例あった．

## 5. 文字サイズの変動

本節では，文字サイズの変動について解析する．サイズ変動が大きい場合，つぶれなどの形状変化も伴うことがある．従って，サイズ変動の程度を観測することで，１カテゴリに必要な標準パターン数の目安にすることができる．

図 4 は，数式領域およびテキスト領域の各文字に生じたサイズ変動のヒストグラムである．ここでは そのカテゴリの標準高さとの比をサイズ変動と定義した．なお，異常文字は除いてある．この図より，数式領域はテキスト領域よりもサイズ変動が

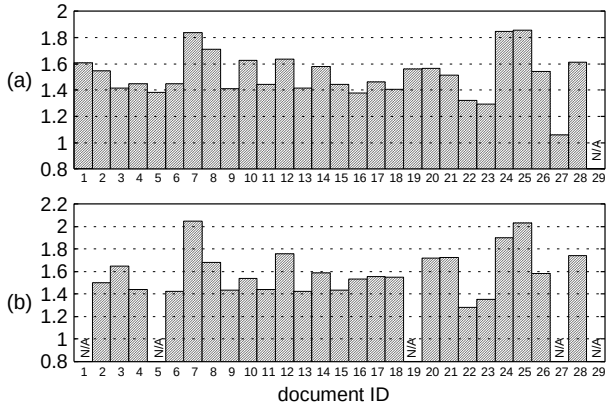


図 5　各文書における，平均ベースライン文字サイズと平均添字サイズの比．(a) アセンダもディセンダも含まないイタリック文字を対象とした場合．(b) イタリック大文字 (“$Q$” を除く) を対象とした場合．該当カテゴリがなかった文書は “N/A” と表記．

Fig. 5　Ratio between the intra-document average size of first-level sub/sup-scripts and that of baseline characters.

大きいことがわかる．具体的には，数式領域においては文字の約 68%以上が平均サイズよりも ±5%以上変動していることがわかる．こうした大きな変動の主たる要因は，括弧や総和記号 (“$\sum$”) のようなサイズ可変の記号類や，添字の存在である．なお，最大変動は 605%の変動，すなわち平均サイズの 6 倍であり，そのカテゴリは “|”(垂直括弧) であった．

数式 OCR では数式構造を解析・認識する際にベースライン文字と添字を区別する必要があるが，サイズ変動を解析することでその困難性を明らかにできる．まず，サイズの逆転（添字がベースライン文字よりも大きくなってしまう場合）を調査した．その結果，ベースライン文字としても添字としても存在した 205 カテゴリのうち，101 カテゴリ，すなわち 49%にこのサイズの逆転が起きていたことがわかった．従って，サイズに関する単純なしきい値処理では両者の区別ができないことがわかる．

次に，各文書について，添字 (2 重や 3 重添字は除く) とベースライン文字のサイズ比について調査した．具体的には，(a) アセンダもディセンダも含まないイタリック小文字 ($a, c, e, m, n, o, \ldots, z$)，ならびに (b) イタリック大文字 ($A, B, C, \ldots, Z$，ただしディセンダを含む $Q$ を除く) について，それぞれベースライン文字と添字での文書内平均サイズの比について調べた．これら (a) や (b) のようにカテゴリを選んだのは，元々ほぼ同じサイズの文字について平均を求めるためである．図 5 はその結果である．全体的には，(a) と (b) のいずれにおいても，ベースライン文字は添字の 1.5 倍ぐらいのサイズであることがわかる．しかしながら，詳細にみると，文献によってかなり異なっていることもわかる．具体的には 2 倍以上サイズの違う文書もあれば，文書 27 に関する (a) の結果のようにほとんど違わない (1.06 倍) 場合もある．この結果から，文書内のサイズ比も，添字判定の基準としては扱いにくいことを示している．以上から，添字判定処理には，各文字の位置関係などサイズ以外の情報が必要と言える．

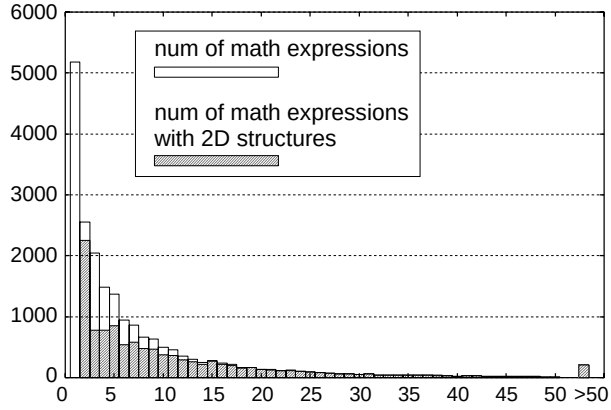


図 6　1 数式あたりの文字数の分布．

Fig. 6　The distribution of the number of characters per math expression.

表 5　リンクの内訳．(上段：個数．下段：割合 (%)．)

Table 5　Contents of links.

| Hori-zontal | Right Sup. | Right Sub. | Left Sup. | Left Sub. | Upper | Under | total |
|---|---|---|---|---|---|---|---|
| 102,573 | 8,706 | 14,767 | 10 | 2 | 1,590 | 4,451 | 132,099 |
| 77.65 | 6.59 | 11.18 | 0.01 | <0.01 | 1.20 | 3.37 | 100.00 |

$$\lim_{r\to\infty} \frac{\sigma_{D''_{w_i}}(r)}{r^{\varrho+2n-2}} = 0\,.$$

図 7　2 重添字と 3 重添字を含む数式．

Fig. 7　High-order sub/sup-scripts.

## 6. 数式の複雑さ

本データベースに含まれる数式の総数は 20,511 であった (インライン数式も含む)．図 6 に 1 数式中の文字数の分布を示す．この結果から，この分布は指数分布で近似できることがわかる．また本図は，数式が時々非常に大規模な (例えば 50 文字以上含む) ものにになることも示している．従って，数式 OCR の数式構造解析処理は，そうした大規模な数式も扱えるよう，十分に実用的な計算量である必要がある．

図 6 に，2 次元構造を持つ数式中の文字数の分布を併せて示す．ここで 2 次元構造を持つ数式とは，水平 (`Horizontal`) 以外のリンクを少なくとも 1 つは持つ数式である．例えば $x^2$ は 2 次元構造を持つが，$a = b + c$ は持たない．この図から，まず数式の多くは 2 次元構造を持つことがわかる．実際，2 文字以上から成る数式の 72%が 2 次元構造を持っていた．また図より，文字数が増えるほど 2 次元構造を持つ割合が増え，15 文字以上からなる数式ではそのほとんどが 2 次元構造を持つことがわかる．以上の結果は，文献 [5], [9], [10] で提案されているような，2 次元構造を意識した数式構造解析処理が重要であることを，定量的に示している．

表 5 に，6 種のリンクの内訳を示す．大半は水平リンクであるが，右添字も 20%弱程度存在することがわかる．左添字はほとんど存在しない．なお，`Upper` の 70.8%，および `Under` の

25.4%は分数線に関するリンクであった．

数式中の添字の総数は 36,901 であった．このうち，34,215 個 (92.72%) が通常の添字 (図 7 の “$D$”，2,615 個 (7.09%) が 2 重添字 (同図 “$w$”)，71 個 (0.19%) が 3 重添字 (同図 “$i$”) であった．4 重以上の添字は存在しなかった．

## 7. ま と め

全体で 453 ページ，約 67 万文字からなる正解付き英語数学文書データベースを定量的に解析することで，数式 OCR 開発における問題点を明らかにすると共に，それらに対する解決案を述べた．主たる解析結果は以下の通りである．

（1） 数式領域，テキスト領域，およびそれらの和である全領域について，それぞれに含まれたカテゴリの数は 320，202，371 であった．後者 2 つを比較することで，数学文書には非数学文書の 2 倍近くのカテゴリが存在することがわかった．

（2） 数式領域，テキスト領域，全領域において，それぞれ約 33%，12%，17%の文字がイタリック (アルファベット) 文字であった．

（3） カテゴリによって発生頻度は大きく異なった．

（4） 数式領域は，テキスト領域に比べて，異常文字の含有率ならびに文字サイズの変動率が高かった．

（5） 異常文字含有率は各文書で大きく異なった．また 1 文書内で特定の異常文字が集中して発生する場合があった．

（6） 数式領域の接触文字の約 20%が，水平以外の方向で接触（添字と非添字の接触）であった．

（7） ベースライン文字のサイズと添字のサイズがほぼ同一となる文書が存在した．

（8） 2 文字以上からなる数式の約 72%が，2 次元構造を有した．特に 15 文字以上からなる数式では，そのほとんどが 2 次元構造を有した．

（9） 数式は時々非常に大きく（50 文字以上）なり，また 3 重添字を含んでいた．ただし 4 重添字は存在しなかった．

## 文 献


[1] K.-F.Chan and D.-Y.Yeung, “Mathematical expression recognition: a survey,” *Int. J. Doc. Anal. Recog.*, 3(1):3-15, 2000.

[2] S.Hara, et al., “MathBraille; a system to transform LATEX documents into Braille,” *SIGCAPH Newsl*, 66:17-20, 2000.

[3] G. O. Michler, “Report on the retrodigitization project “Archiv der Mathematik”,” *Archiv der Mathematik*, 77:116-128, 2001.

[4] K.Dennis, G.O.Michler, G.Schneider, and M.Suzuki, “Automatic reference linking in distributed digital libraries,” *Proc. Workshop Doc. Image Anal. and Retrieval (DIAR-03)*, 2003.

[5] H.-J. Lee and J.-S. Wang, “Design of a mathematical expression understanding system,” *Pattern Recognition Letters*, 18(3):289–298, 1997.

[6] M. Okamoto, H. Imai, and K. Takagi, “Performance evaluation of a robust method for mathematical expression recognition,” *Proc. ICDAR*, 121-128, 2001.

[7] M. Okamoto, S. Sakaguchi, and T. Suzuki, “Segmentation of touching characters in formulas,” in *Doc. Anal. Sys.: Theory and Practice. Third IAPR Workshop, DAS’98. Selected Papers* (Lect. Note in Comput. Sci. vol.1655), Springer-Verlag, 1999.

[8] A. Nomura, K. Michishita, S. Uchida, and M. Suzuki, “Detection and segmentation of touching characters in mathematical expressions,” *Proc. ICDAR*, 1:126-130, 2003.

[9] J. Ha, R. M. Haralick, and I. T. Phillips, “Understanding mathematical expressions from document images,” *Proc. ICDAR*, 956-959, 1995.

[10] Y. Eto and M. Suzuki, “Mathematical formula recognition using virtual link network,” *Proc. ICDAR*, 762–767, 2001.


## 付 録

### 1. データベース中の文献リスト

本論文で対象とした数学文書データベースに含まれる英語論文 29 編のリストを以下に示す．いずれも (純粋) 数学に関するものである．


● J.A.Jenkins and K.Oikawa, “On the growth of slowly increasing unbounded harmonic functions,” Acta Math., 124(1-2), 37-63, 1970. ● A.Mcdaniel and L.Smolinsky, “Lax equations,weight lattices, and Prym-Tjurin varieties,” *ibid.*, 181(2), 283-305, 1998. ● J.S.Milne, “Weil-Chatelet groups over local fields,” Ann. Sci. Ecole Norm. Sup., 4d sér, t.3, 273-284, 1970. ● S.Nollet, “The Hilbert schemes of degree three curves,” *ibid.*, t.30, 367-384, 1997. ● J.-E.Bjork, “Every compact set in $C^N$ is a good compact set,” Ann. Inst. Fourier, 20(1), 493-498, 1970. ● M.Suzuki, “Affine plane curves with one place at infinity,” *ibid.*, 49(2), 375-404, 1999. ● R.Osserman, “A proof of the regularity everywhere of the classical solution to Plateau’s problem,” Ann. Math., 91, 550-569, 1970. ● F.Gardiner, “Deformations of embeddings of Riemann surfaces in projective space, Advances in the theory of Riemann surfaces, “ Ann. Math. Studies, 66, 157–173, 1971. ● L.Gruman, “Entire functions of several variables and their asymptotic growth,” Arkiv für Matematik, 9(1), 141-163 1971. ● H.-M.Maire and F.Meylan, “Extension of smooth CR mappings between non-essentially finite hypersurfaces in $C^3$,” *ibid.*, 35(1), 185-199, 1997. ● B.-Y.Chen, “On an inequality of mean curvatures of higher degree,” Bull. Amer. Math. Soc., 77(1), 157-163 1971. ● L.Bers, “On spaces of Riemann surfaces with nodes,” *ibid.*, 80(6), 1219-1222, 1974. ● F.Morel, “Voevodsky’s proof of Milnor’s conjecture,” *ibid.*, 35(2), 123-143, 1998. ● P.J.Grabe and G.Viljoen, “Maximal classes of ext-reproduced Abelian groups,” Bull. Soc. Math. France, 98, 165-192, 1970. ● M.Andresson, J.Boo, and J.Ortega-Cerda, “Canonical homotopy operators for the $\bar{\delta}$ vomplex in strictly pseudo-convex domains,” *ibid.*, 126, 245-271, 1998. ● M.Hirsch, J.Palis, C.Pugh, and M.Shub, “Neighborhoods of hyperbolic sets,” Invent. Math., 9, 121-134, 1970. ● K.Altmann and J.Stevens, “Cotangent cohomology of rational surface singularities,” *ibid.*, 138, 163-181, 1999. ● A.Kono, “On cohomology mod 2 of the classifying spaces of non-simply connected classical Lie groups,” J. Math. Soc. Japan, 27(2), 281-288, 1975. ● M.Ishida, “On the genus field of an algebraic number field of odd prime degree,” *ibid.*, 27(2), 289-293, 1975. ● T.Itoh, “On Veronese manifold,” *ibid.*, 27(2), 497-506, 1975. ● Y.Kusunoki and Y.Sainouchi, “Holomorphic differentials on open Riemann surfaces,” J. Math. Kyoto Univ., 11(1), 181-194, 1971. ● S.Suzuki, “On Neggers’ numbers of discrete valuation rings,” *ibid.*, 11(1), 373-375, 1971. ● S.Suzuki, “Differential modules and derivations of complete discrete valuation rings,” *ibid.*, 11(2), 377-379, 1971. ● T.Hara, “Equivariant SK invariants on $\mathbb{Z}_{2^r}$ manifolds with boundary,” Kyushu J. Math., 53, 17-36, 1999. ● K.Diederich, “Pseudoconvex domains: an example with nontrivial nebenhulle,” Math. Ann., 225(3), 275-292, 1977. ● G.Cornelissen, “Zeros of Eisenstein series, quadratic class numbers and supersingularity for rational function fields,” *ibid.*, 315, 175-196, 1999. ● K.Kitano, “The growth of the resolvent and hyperinvariant subspaces,” Tohoku Math. J., 25, 317-331, 1973. ● K.Saka, “A note on subalgebras of a measure algebra vanishing on non-symmetric homomorphisms,” *ibid.*, 25, 333-338, 1973. ● M.Muro, “Invariant hyperfunctions on regular prehomogeneous vector spaces of commutative parabolic type,” *ibid.*, 42, 163-193, 1990.