A Ground-Truthed Mathematical Character and Symbol Image Database

Masakazu Suzuki*, Seiichi Uchida** and Akihiro Nomura***
* Faculty of Mathematics,
** Faculty of Information Science and Electrical Engineering,
*** Graduate School of Mathematics,
Kyushu University, 6-10-1 Hakozaki, Higashi-ku, Fukuoka-shi, 812-8581 Japan

Abstract

This paper describes the specifications for our ground-truthed mathematical character and symbol image database, called InftyCDB-1. The ground-truth of each character is composed of type, font, quality (touched/broken) and link (relative position), etc. The database includes all the characters and symbols of 467 pages of 30 articles on mathematics, and is organized so that it can be used as word image database or as mathematical formula image database. InftyCDB-1 is a public database that is freely usable for research and development purposes.

1 Introduction

In this paper, we report the specifications for our ground-truthed mathematical character and symbol image database, called InftyCDB-1, which is freely usable for research and development purposes. The ground-truth of each character is composed of type, font, quality (touched/broken) and link (relative position), etc. The database includes all the characters and symbols of 467 pages of 30 articles on mathematics, and is organized so that it can be used as a word image database or as a mathematical formula image database. Thus, the database can be used, for example, in the following ways for researches:

- development and evaluation of character and scientific symbol recognition,
- development and evaluation of mathematical formula recognition,
- analysis of words in mathematical documents.

Since all the character images that appear in the page images are included in the database, users can get training data or test data for character/symbol recognition. For all the special mathematical symbols in the database, their own code and symbol name have been carefully attached. Since each alpha-numeric character in the database has its font attributes such as italic/upright, bold or not, the database can be used for evaluations of font distinction ability in character recognition.

The image data are stored separated into word or math formula units and arranged in alphabetic order independent of the content of papers. No whole page image is included in the database to avoid copyright problems.

Hereafter, the term *character* means not only ordinary characters (e.g., "A"), but also math symbols (e.g., "+"), unless otherwise noted. The term *category* means the finest level of character classification and the term *type* means a set of categories having a similar property. For example, "A", "B" and "C" are three categories belonging to the same type (Roman). In contrast, "A"(Roman), "A"(italic), "$\mathcal{A}$"(calligraph), "$\mathbb{A}$"(blackboard bold), "$\mathfrak{A}$" (German), and "$\mathcal{A}$" (script) are six categories belonging to different types. Each character belongs to either the *text region* or the *math region*. The math region includes not only numbered equations but also in-line math formulae. Note that many in-line math formulae are composed of a single character, such as "x" in the sentence "The variable x denotes ...".

2 Outline of database

2.1 Data collection

The documents contained in the database are 30 English articles on pure mathematics (published 1970 $\sim$ 2000). The numbers of pages, characters, words and math expressions in the database are 467, 688,570, 108,914 and 21,056, respectively. For a quantitative analysis of the database, see [1][1] . This database is larger than past databases for research on math-OCR (e.g., about 15,000 characters in [2], about 10,000 characters in [3]). Note that matrices, tables, and figures are excluded from the database.

All pages were scanned in 600 dpi and binarized automatically by the same commercial scanner (RICOH Imagio Neo 450). The quality of the resulting page images varies

[1]There are slight differences between the table1 below and the table in [1] because some errors were found after submission of the paper [1] and corrected

Table 1. Characters in the database.

type	font	category examples	#predefined categories	text region #cat.	text region #char (%)	math region #cat.	math region #char (%)	total #cat.	total #char (%)
accent		ˆ ˜ ¯ ¨ ˋ ˇ _	13	1	2 (<0.01)	7	2,700 (1.72)	7	2,702 (0.39)
arrow		← ↔ ↖	16	1	3 (<0.01)	7	1,103 (0.70)	7	1,106 (0.16)
big symbol		∑ ∫ ∏	18	0	0 (0.00)	11	2,458 (1.57)	11	2,458 (0.36)
blackboard bold		𝔸 𝔹 ℂ 𝔻 𝔼 𝔽	26	0	0 (0.00)	9	427 (0.27)	9	427 (0.06)
calligraphic		𝒜 ℬ 𝒞 𝒟 ℰ ℱ	26	0	0 (0.00)	19	592 (0.38)	19	592 (0.09)
German	Upright	𝔄 𝔅 ℭ 𝔞 𝔟 𝔠	52	0	0 (0.00)	25	1,041 (0.66)	25	1,041 (0.15)
	Bold	𝖀 𝖁 𝖂 𝖆 𝖇 𝖈	52	0	0 (0.00)	0	0 (0.00)	0	0 (0.00)
Greek	Upright	Γ Δ Θ	11	0	0 (0.00)	10	2,148 (1.37)	10	2,148 (0.31)
	Italic	$\alpha\beta\gamma$	29	5	19 (<0.01)	23	10,618 (6.76)	23	10,637 (1.54)
	Bold	**Γ Δ Θ**	11	0	0 (0.00)	1	3 (<0.01)	1	3 (<0.01)
	Italic Bold	$\boldsymbol{\alpha\beta\gamma}$	29	0	0 (0.00)	5	31 (0.02)	5	31 (<0.01)
extended Latin	Upright	Ä Æ è	182	30	392 (0.07)	2	3 (<0.01)	30	395 (0.06)
	Italic	*Ä Æè*	182	9	55 (0.01)	2	10 (0.01)	10	65 (0.01)
	Bold	**Ä Æ è**	182	4	6 (<0.01)	0	0 (0.00)	4	6 (<0.01)
	Italic Bold	***Ä Æè***	182	0	0 (0.00)	0	0 (0.00)	0	0 (0.00)
numeric	Upright	0 1 2	10	10	12,018 (2.26)	10	15,294 (9.74)	10	27,312 (3.97)
	Italic	*0 1 2*	10	10	140 (0.03)	4	118 (0.08)	10	258 (0.04)
	Bold	**0 1 2**	10	10	923 (0.17)	4	26 (0.02)	10	949 (0.14)
	Italic Bold	***0 1 2***	10	0	0 (0)	0	0 (0.00)	0	0 (0.00)
operator		+ − × / < &	92	6	154 (0.03)	49	20,359 (12.96)	50	20,513 (2.98)
others	Upright	§ @ © ∞ ∀ ∃ ∂	42	10	2,903 (0.55)	15	1,797 (1.14)	20	4,700 (0.68)
	Bold	**§ @ ©**	16	3	42 (0.01)	0	0 (0.00)	3	42 (0.01)
parenthesis	Upright	( ) { } [ ]	20	7	8,082 (1.52)	12	30,334 (19.31)	12	38,416 (5.58)
	Bold	**( ) { } [ ]**	20	2	112 (0.02)	0	0 (0.00)	2	112 (0.02)
point	Upright	, . ‘ ’	17	11	21,599 (4.06)	11	8,443 (5.41)	14	30,042 (4.36)
	Bold	**, . ‘ ’**	17	6	469 (0.09)	0	0 (0.00)	6	469 (0.07)
Roman	Upright	A B C a b c	61	57	414,825 (78.05)	55	8,259 (5.26)	57	423,084 (61.44)
	Italic	*A B C a b c*	61	55	63,590 (11.96)	53	49,072 (31.24)	56	112,662 (16.36)
	Bold	**A B C a b c**	61	56	6,178 (1.16)	13	538 (0.34)	56	6,716 (0.98)
	Italic Bold	***A B C a b c***	61	0	0 (0.00)	19	1,508 (0.96)	19	1,508 (0.22)
script		𝒜 ℬ 𝒞	52	0	0 (0.00)	7	176 (0.11)	7	176 (0.03)
total			1,571	294	531,512 (100.00)	373	157,058 (100.00)	487	688,570 (100.00)

Notes: (1) Each “Roman” and “italic” type includes nine double letters (i.e., ligatures), such as “fi”.

with the quality of original print and/or copy. Several page images are noisy and include a lot of abnormal characters, such as touching or broken characters.

2.2 Ground truth

The ground truth for each character was attached *manually* by seven students in, or a graduate from, a university math department. The ground truth of each character is composed of the following attributes:

- type, category and font
- text or math region
- normal or abnormal character
- size (height and width)
- link
- location in the word or formula image
- path (folder name + file name) to the image file.

The fifth attribute, link, represents the positional relationship to the preceding character and was attached to describe the structure of a math formula (as a tree). There

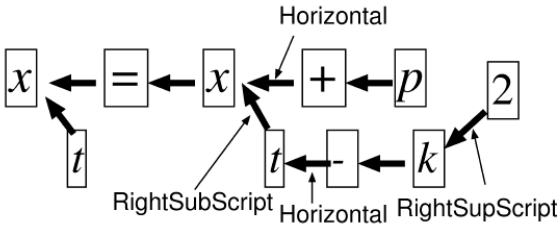

Figure 1. Link example of a math expression.

are six kinds of links: horizontal, right-superscript, right-subscript, left-superscript, left-subscript, upper, and lower. Figure 1 shows the link structure of a formula $x_t = x_{t-k^2} + p$. The sixth attribute, location, is the rectangular coordinates (left,top,right,bottom) of the character image in the corresponding word/formula image. Each character data is connected to the word/formula image data by the seventh attribute, path, and the sixth attribute, location.

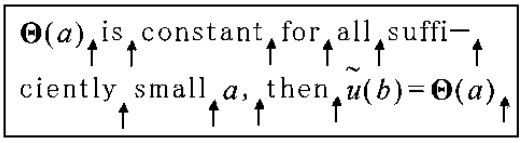

Figure 2. Word/formula segmentation.

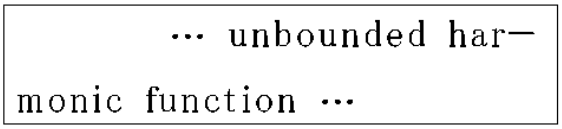

Figure 3. SyllableAfter attribute for hyphenation.

2.3 Word segmentation

The segmentation of words is basically done by spacing. On the other hand, a set of consecutive math expressions in a line are unified into one formula regardless of the spacing, e.g., before and after a relative operators such as equal sign, etc. A word or a formula continued to next line is separated at the end of a line. Points (e.g. ",", ".", etc) are included in the word/formula just before the points. Opening parenthesis are included in the next word and closing parenthesis are included in the preceding words like points. Quotation marks are treated in a similar way.

In Figure 2, the arrows show the segmentation points and the sentence is separated into 11 words/formulae in this example.

3 Structure of database

The database InftyCDB-1 is composed of two parts: (i) text data and (ii) image data, related to each other. Text data is a Microsoft Access or CSV-format, the user's choice, while image data are systematically named PNG files.

3.1 Text data

For each character, the 29 attributes listed in the Table 2 are attached:

The attribute (5) is code defined in our laboratory to distinguish character/symbol categories in the math-OCR software called InftyReader[5]. The attribute (6) is a string to read the character: e.g., "int" for "$\int$", "Omega" for "Ω". The attribute (7) is "text" for text region character, and "math" for math region character. The attribute (8) is "True" (resp. "False") if the character is on the baseline (resp. in sub/super-script area). The attribute (9) (resp. (10) is "True" if the character is italic (resp. bold) font and "False" otherwise. The attribute (11) is "touched" for a touched character, "separate" for a broken character, "touch_ and_ sep" for touched and broken character, and "normal" otherwise. By using the attributes (14), (15), a user can reproduce the math tree structure for each formula.

As for the rule to define the path to the image file in the attribute (16), see 3.2 below.

Attributes (21) – (29) are the same for all the characters in a word/formula. Attribute (21) is the ID number attached to each word/formula. Attributes (22), (23) and (24) are string data that represent the corresponding word/formula in MathML format, LaTeX format, and IML format, respectively. IML format is the XML format used in the software developed in our laboratory, math-OCR "InftyReader" and an authoring tool "InftyEditor" for mathematical documents. The attribute (29) is "True" for words at the beginning of line continued by hyphenation from the word at the end of previous line, and "False" for other cases. In Fifure 3, the latter part "monic" of "harmonic" has the SyllableAfter attribute "True".

A sample of the text data for a formula

$$\frac{\mathrm{d}}{\mathrm{dt}}\mathrm{h}_{\mathrm{t}\nu,\mathrm{z}_0}|_{\mathrm{t}=0}$$

and a word "and" is shown in Table 3 below.

3.2 Image data

To reduce the number of image files, the images of a same word in a same article are stored in one image file. However, italic words and upright words are stored in different image files, and Roman capital/small letter are distinguished so that, e.g., "And" "and" are stored in different image files. Formula images in an article are also grouped in a same way, when the expressions are identical. Image file names are defined as follows:

- word $\cdots$ "string"(_ FontFlag)_ "number".png
- formula $\cdots$ "number of characters in the formula" _"first three characters in the formula" (_ FontFlag)_ "number".png

For example, the file names of the word "(and)" and the formula "$\alpha \leq 1$" are "LeftPar-and-RightPar_0.png" and "3_alpha-le-1_0.png", respectively. Figure 4 shows two image files in the database having the path names "Images¥ActaM_1970_37_63¥A¥approaches_I_0.png" and "Images¥ActaM_1970_37_63¥MATH¥1-9¥7_D-Omega-BigLeftPar_0.png", respectively.

4 Distribution

The database InftyCDB-1 is made and will be maintained in M.Suzuki laboratory, Faculty of Mathematics, Kyushu University. It is freely available for research and development purposes after user registration. (http://www.inftyproject.org/).

Table 2. List of attributes.

	Attribute	Explanation
(1)	CharID	ID number of each character
(2)	JornalID	ID number of the article
(3)	SheetID	Page number
(4)	Type	Type name (see Table 1)
(5)	Code	Category code (OCR code)
(6)	Entity	Category name
(7)	Region	Distinction of text/math area
(8)	Baseline	Distinction of baseline/subscript
(9)	ItalicFlag	Italic flag
(10)	BoldFlag	Bold flag
(11)	Quality	Normal/touched/separate/touch_ and_ sep
(12)	Width	Width of the character
(13)	Height	Height of the character
(14)	ParentCharID	Parent CharID of the link
(15)	Link	Link name
(16)	ImageName	Path (folder name + file name) to the image file including the character
(17)(18)(19)(20)	Rect	Coordinates of the character in the image file (left,top,right,bottom)
(21)	WordID	ID number of the word/formula including the cahracter
(22)	WordMathML	MathML string
(23)	WordTeX	LaTeX string
(24)	WordIML	IML string
(25)(26)(27)(28)	WordRect	Coordinates of the word/formula in the image file (left,top,right,bottom)
(29)	SyllableAfter	Flag of word continued from the prevoious line by hyphenation

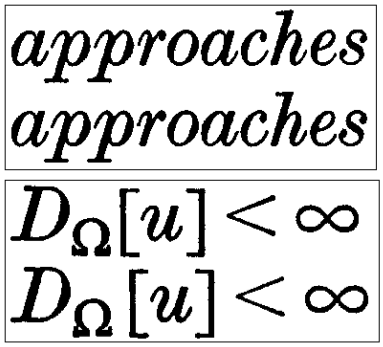

Figure 4. Example of image files

The text data part is in CSV (16.8MB) or Microsoft Access format (26.7MB) as the user chooses, and the image data is in PNG format (202MB). The database is delivered in CD-ROM.

5 Conclusion

In this paper, we described our ground-truthed mathematical character and symbol image database, called InftyCDB-1. The database consists of two parts: text data and image data that are related to each other.

The ground-truth of each character is composed of type, font, quality (touched/broken) and link information to represent the tree structure of math formula, etc. The database includes all the 688,570 characters (and symbols) of 467 pages of 30 English articles on mathematics (published 1970∼ 2000). Characters are grouped into words/formulae in the database. Total number of words and formulae in the database are 108,914 and 21,056, respectively. The database is freely usable for research, development and evaluation of math-OCRs.

References

[1] S. Uchida, A. Nomura, M. Suzuki "Quantitative Analysis of Mathematical Documents," *Int. J. Doc. Anal. Recog*., to appear.

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

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

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

[5] M. Suzuki, F. Tamari, R. Fukuda, S. Uchida, T. Kanahori "Infty- an integrated OCR system for mathematical documents," *ACM Symposium on Document Engineering*, 95-104, 2003

Table 3. Example of the text data in the database for a formula "$\frac{\mathrm{d}}{\mathrm{dt}}\mathrm{h}_{\mathrm{t}\nu,\mathrm{z}_0}|_{\mathrm{t}=0}$" and a word "and".

(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)	(12)	(13)
14	5	10	BigSymbol	33D1	fractionalLine	math	TRUE	FALSE	FALSE	normal	82	12
15	5	10	Roman	0164	d	math	TRUE	FALSE	FALSE	normal	38	60
16	5	10	Roman	0164	d	math	TRUE	FALSE	FALSE	normal	38	59
17	5	10	Roman	0174	t	math	TRUE	FALSE	FALSE	normal	24	53
18	5	10	Roman	0168	h	math	TRUE	FALSE	FALSE	normal	39	60
19	5	10	Roman	0174	t	math	FALSE	FALSE	FALSE	normal	27	52
20	5	10	Greek	426D	nu	math	FALSE	TRUE	FALSE	normal	42	37
21	5	10	Point	142C	comma	math	FALSE	FALSE	FALSE	normal	17	26
22	5	10	Roman	017A	z	math	FALSE	FALSE	FALSE	normal	38	39
23	5	10	Numeric	0130	zero	math	FALSE	FALSE	FALSE	normal	34	49
24	5	10	Parenthesis	197C	vert	math	TRUE	FALSE	FALSE	normal	11	187
25	5	10	Roman	0174	t	math	FALSE	FALSE	FALSE	normal	26	52
26	5	10	Operator	1D3D	equal	math	FALSE	FALSE	FALSE	touched	43	22
27	5	10	Numeric	0130	zero	math	FALSE	FALSE	FALSE	normal	35	48
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮
804	2	13	Roman	0161	a	TRUE	text	FALSE	FALSE	normal	37	38
805	2	13	Roman	016E	n	TRUE	text	FALSE	FALSE	separate	42	36
806	2	13	Roman	0164	d	TRUE	text	FALSE	FALSE	normal	42	56

• • •

(14)	(15)	(16)	(17)	(18)	(19)	(20)	(21)	(22)	(23)	(24)	(25)	(26)	(27)	(28)	(29)
-1	TOP	∗1	0	73	82	85	28005695	∗2	∗3	∗4	0	0	522	205	FALSE
14	UNDER	∗1	12	101	50	161	28005695	∗2	∗3	∗4	0	0	522	205	FALSE
14	UPPER	∗1	24	0	62	59	28005695	∗2	∗3	∗4	0	0	522	205	FALSE
15	HORIZONTAL	∗1	54	109	78	162	28005695	∗2	∗3	∗4	0	0	522	205	FALSE
14	HORIZONTAL	∗1	95	35	134	95	28005695	∗2	∗3	∗4	0	0	522	205	FALSE
18	RSUB	∗1	135	78	162	130	28005695	∗2	∗3	∗4	0	0	522	205	FALSE
19	HORIZONTAL	∗1	173	96	215	133	28005695	∗2	∗3	∗4	0	0	522	205	FALSE
20	HORIZONTAL	∗1	234	119	251	145	28005695	∗2	∗3	∗4	0	0	522	205	FALSE
21	HORIZONTAL	∗1	262	92	300	131	28005695	∗2	∗3	∗4	0	0	522	205	FALSE
22	RSUB	∗1	306	116	340	165	28005695	∗2	∗3	∗4	0	0	522	205	FALSE
18	HORIZONTAL	∗1	363	18	374	205	28005695	∗2	∗3	∗4	0	0	522	205	FALSE
24	RSUB	∗1	386	130	412	182	28005695	∗2	∗3	∗4	0	0	522	205	FALSE
25	HORIZONTAL	∗1	423	154	466	176	28005695	∗2	∗3	∗4	0	0	522	205	FALSE
26	HORIZONTAL	∗1	487	135	522	183	28005695	∗2	∗3	∗4	0	0	522	205	FALSE
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮
-1	TOP	∗5	0	5850	37	5888	1000299	and	and	and	0	5831	129	5888	FALSE
804	HORIZONTAL	∗5	40	5851	82	5887	1000299	and	and	and	0	5831	129	5888	FALSE
805	HORIZONTAL	∗5	87	5831	129	5887	1000299	and	and	and	0	5831	129	5888	FALSE

(∗1) AnnMS_1971_157_173¥MATH¥10-19¥14_fractionalLine-dd_0.png, (∗5) ActaM_1970_37_63¥A¥and_1.png

(∗2),(∗3), (∗4): Expression of the formula $\frac{\mathrm{d}}{\mathrm{dt}}\mathrm{h}_{\mathrm{t}\nu,\mathrm{z}_0}|_{\mathrm{t}=0}$ in MathML, in LaTeX and in IML respectively.

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

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

List of articles in the database

• Acta Math., 124(1-2), 37-63, 1970. • *ibid.*, 181(2), 283-305, 1998. • Ann. Sci. Ecole Norm. Sup., 4d sér, t.3, 273-284, 1970. • *ibid.*, t.30, 367-384, 1997. • Ann. Inst. Fourier, 20(1), 493-498, 1970. • *ibid.*, 49(2), 375-404, 1999. • Ann. Math., 91, 550-569, 1970. • Ann. Math. Studies, 66, 157–173, 1971. • Arkiv für Matematik, 9(1), 141-163 1971. • *ibid.*, 35(1), 185-199, 1997. • Bull. Amer. Math. Soc., 77(1), 157-159 1971. • *ibid.*, 77(1), 160-163 1971. • *ibid.*, 80(6), 1219-1222, 1974. • *ibid.*, 35(2), 123-143, 1998. • Bull. Soc. Math. France, 98, 165-192, 1970. • *ibid.*, 126, 245-271, 1998. • Invent. Math., 9, 121-134, 1970. • *ibid.*, 138, 163-181, 1999. • J. Math. Soc. Japan, 27(2), 281-288, 1975. • *ibid.*, 27(2), 289-293, 1975. • *ibid.*, 27(2), 497-506, 1975. • J. Math. Kyoto Univ., 11(1), 181-194, 1971. • *ibid.*, 11(1), 373-375, 1971. • *ibid.*, 11(2), 377-379, 1971. • Kyushu J. Math., 53, 17-36, 1999. • Math. Ann., 225(3), 275-292, 1977. • *ibid.*, 315, 175-196, 1999. • Tohoku Math. J., 25, 317-331, 1973. • *ibid.*, 25, 333-338, 1973. • *ibid.*, 42, 163-193, 1990.