

From Paper to XML in Mathematics

April 26, 2004

IMA workshop "Enhancing the Searching of Mathematics"
Univ. Minnesota, Minneapolis

Masakazu Suzuki
Kyushu University

Suzuki@math.kyushu-u.ac.jp

Section 1

Introduction

Motivation

■ Digitization of mathematical Journals

Searchable by Keywords, Definitions, Theorems, etc.

■ Re-usability of data

Reproduction of old books,
Conversion to LaTeX source or XML data base,
Verification by computer algebra systems,
Knowledge database of mathematical theorems, etc.

■ Automatic transcription

Transcription into other languages, into Braille codes, etc.

Different levels in digitization

- Level 1: Bitmap images of printed materials
e.g. GIF, TIFF
- Level 2: Searchable digitized document
e.g. PDF with hidden text
- Level 3: Structured document with links
e.g. XML, HTML(+MathML), LATEX, ...
- Level 4: (partially) Executable document
e.g. Mathematica, Maple
- Level 5: Formally presented document.
e.g. Mizar, OMDoc

- We released a new version on 15/03/2004:

InftyReader Ver.2.4.1

downloadable from our web site:

<http://infty.math.kyushu-u.ac.jp>


- Two editions:

1. *InftyReaderE*, English edition, free, no time limit, uses no commercial OCR engine.
2. *InftyReaderJ*, Japanese edition, free, time limited, uses Toshiba's ExpressReader SDK for Japanese character recognition.

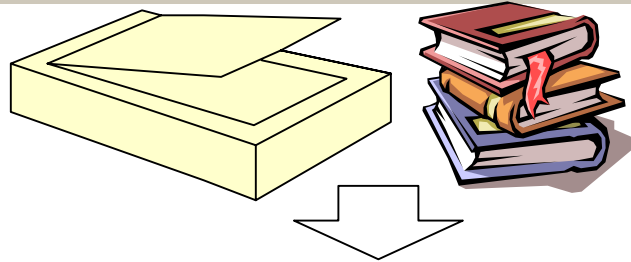
Different levels in digitization

- Level 1: Bitmap images of printed materials
e.g. GIF, TIFF
- Level 2: Searchable digitized document
e.g. PDF with hidden text
- Level 3: Structured document with links
e.g. XML, HTML(+MathML), LATEX, ...
- Level 4: (partially) Executable document
e.g. Mathematica, Maple
- Level 5: Formally presented document.
e.g. Mizar, OMDoc

Different levels in digitization

- Level 1: Bitmap images of printed materials
e.g. GIF, TIFF
- Level 2:  InftyReader : Level 1 Level 3
e.g. PDF w
- Level 3: Structured document with links
e.g. XML, HTML(+MathML), LATEX, ...
- Level 4: (partially) Executable document
e.g. Mathematica, Maple
- Level 5: Formally presented document.
e.g. Mizar, OMDoc

INFTY's Flow

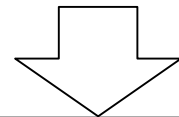


Ideal Flow

Segmentation of Areas (Text, Table, Figure)

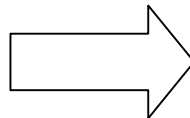
Recognition per line
(Character recognition, Math. Structure analysis)

Document Structure analysis
(Chapter, Section, Itemize, Theorem description, References, etc.)



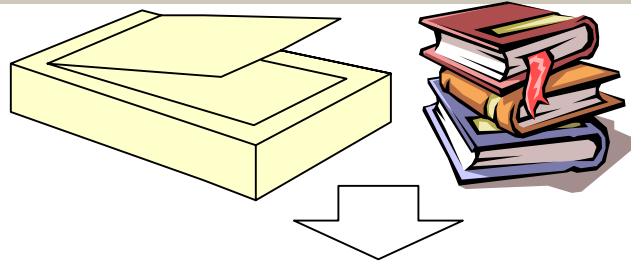
XML

Outputs



LaTeX, HTML+MathML,
PDF, Braille codes, etc.

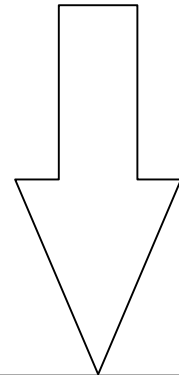
INFTY's Flow



Current Version

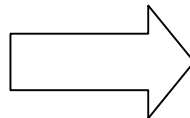
Segmentation of Areas (Text, Table, Figure)

Recognition per line
(Character recognition, Math. Structure analysis)



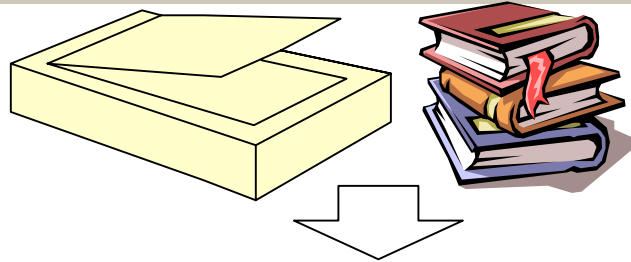
XML

Outputs



LaTeX, HTML+MathML,
PDF, Braille codes, etc.

INFTY's Flow

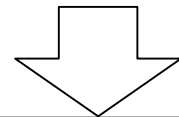


Future Version
(next version?)

Segmentation of Areas (Text, Table, Figure)

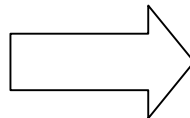
Recognition per line
(Character recognition, Math. Structure analysis)

Document Structure analysis
(Chapter, Section, Itemize, Theorem description, etc.)



XML

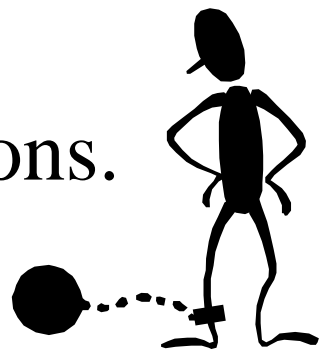
Outputs



LaTeX, HTML+MathML,
PDF, Braille codes, etc.

Difficulty of Math. recognition

- Symbols (Greeks, various math. symbols...)
- Fonts (Italic, Bold, Bbb, Caligraphic, etc.)
- Variation of sizes (subscripts, big integral, big summation symbol, etc.)
- From two dimensional layout structure to mathematical context
- No “word dictionary” in math. expressions.
- Distinction of noises and small symbols

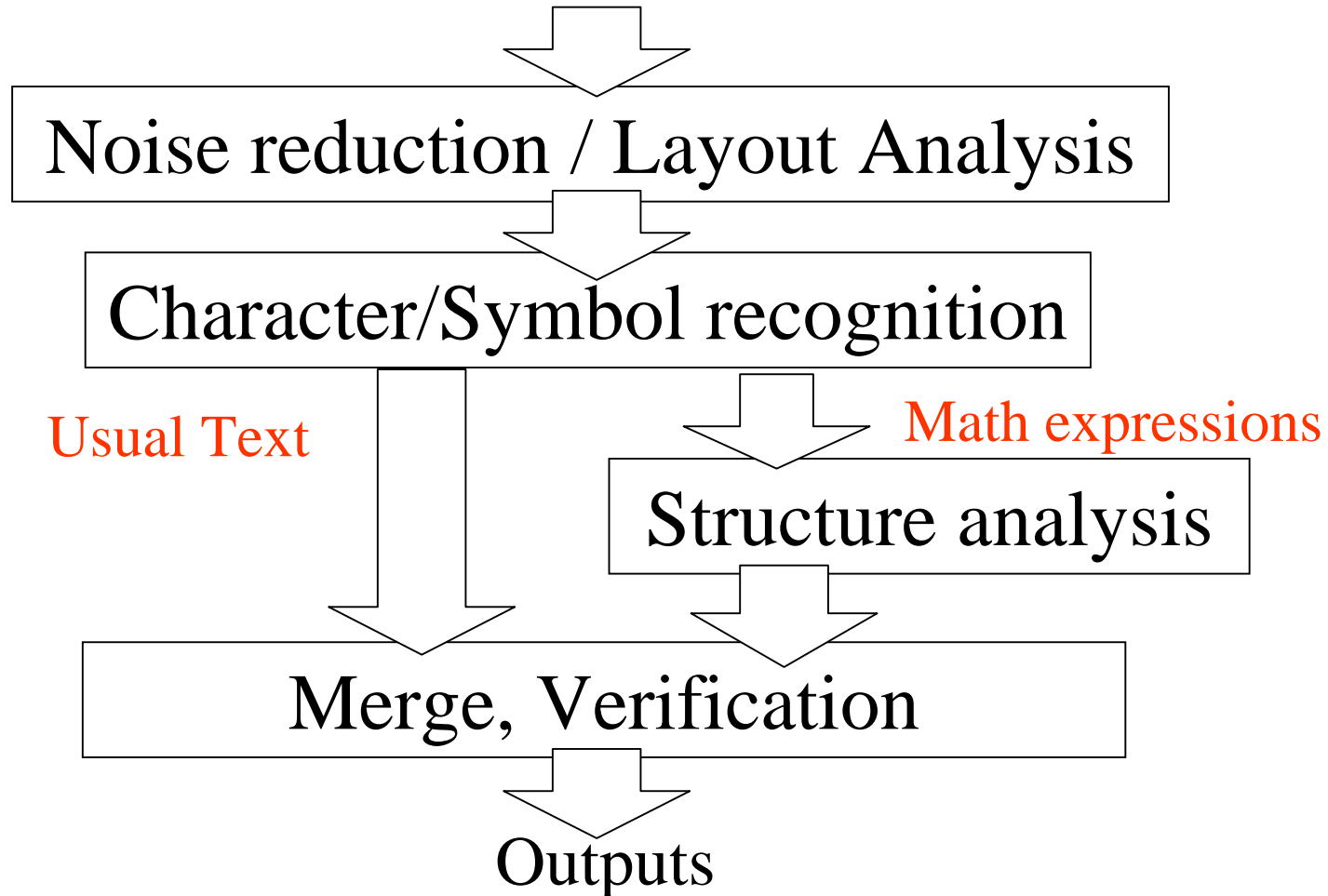
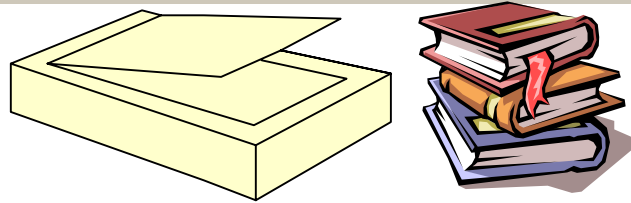


- Recognize page images of printed mathematical documents scanned by 400/600dpi,
- Automatic segmentation of usual text parts, math-expression parts, figures, tables, etc.,
- Recognition of mathematical formulae structures,
- Outputs into various formats: LaTeX, XML, MathML, Braille codes, ...

- We have two styles of applications:
 1. *InftyReader* (free software version) downloadable from our web site: <http://infty.math.kyushu-u.ac.jp>
 2. *InftyReader Pro* (professional version), not yet open to public.

- Demonstration....

INFTY's Recognition Flow



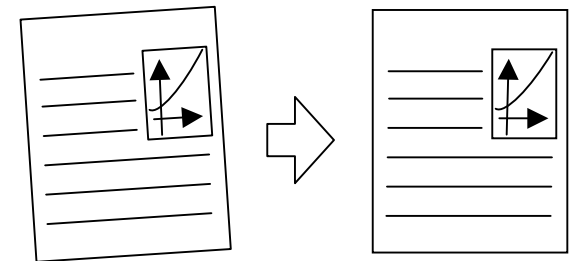
Section 2

Layout Analysis

Process of layout analysis

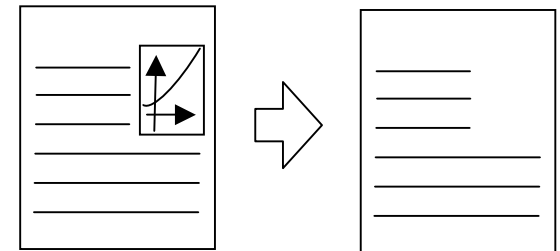
- Pre-processing

- Noise reduction
- Skew correction

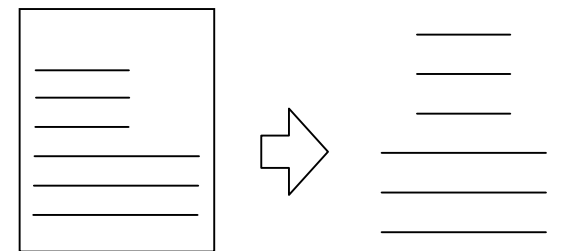


- Segmentation of Figures and Tables

- Analysis of big symbols and small dot patterns



- Segmentation of Math/Text areas into lines



Layout Samples

4

1 The Complex Numbers

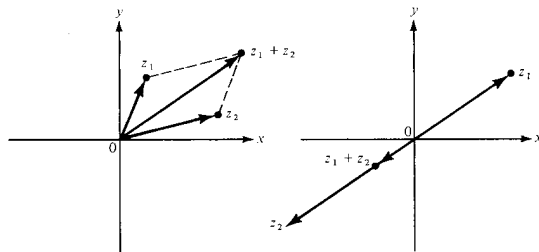
In Chapter 5, we will see that quadratic equations are not unique in this respect: every nonconstant polynomial with complex coefficients has a zero in the complex field.

One property of real numbers that does not carry over to the complex plane is the notion of *order*. We leave it as an exercise for those readers familiar with the axioms of order to check that the number i cannot be designated as either positive or negative without producing a contradiction.

1.2 The Complex Plane

Thinking of complex numbers as ordered pairs of real numbers (a, b) is closely linked with the geometric interpretation of the complex field, discovered by Wallis, and later developed by Argand and by Gauss. To each complex number $a + bi$ we simply associate the point (a, b) in the Cartesian plane. Real numbers are thus associated with points on the x -axis, called the *real axis* while the purely imaginary numbers bi correspond to points on the y -axis, designated as the *imaginary axis*.

Addition and multiplication can also be given a geometric interpretation. The sum of z_1 and z_2 corresponds to the vector sum: If the vector from 0 to z_2 is shifted parallel to the x and y axes so that its initial point is z_1 , the resulting terminal point is $z_1 + z_2$. If 0, z_1 and z_2 are not collinear this is the so-called parallelogram law; see below.



The geometric method for obtaining the product $z_1 z_2$ is somewhat more complicated. If we form a triangle with two sides given by the vectors (originating from 0 to) 1 and z_1 , and then form a similar triangle with the same orientation and the vector z_2 corresponding to the vector 1, the vector which then corresponds to z_1 will be $z_1 z_2$.

This can be verified geometrically but will be most transparent when we introduce polar coordinates later in this section. For the moment, we



1

The Complex Numbers

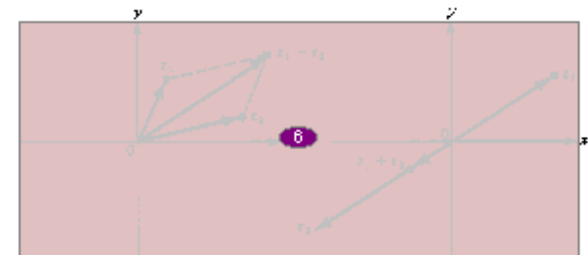
In Chapter 5, we will see that quadratic equations are not unique in this respect: every nonconstant polynomial with complex coefficients has a zero in the complex field.

One property of real numbers that does not carry over to the complex plane is the notion of *order*. We leave it as an exercise for those readers familiar with the axioms of order to check that the number i cannot be designated as either positive or negative without producing a contradiction.

1.2 The Complex Plane

Thinking of complex numbers as ordered pairs of real numbers (a, b) is closely linked with the geometric interpretation of the complex field, discovered by Wallis, and later developed by Argand and by Gauss. To each complex number $a + bi$ we simply associate the point (a, b) in the Cartesian plane. Real numbers are thus associated with points on the x -axis, called the *real axis* while the purely imaginary numbers bi correspond to points on the y -axis, designated as the *imaginary axis*.

Addition and multiplication can also be given a geometric interpretation. The sum of z_1 and z_2 corresponds to the vector sum: If the vector from 0 to z_2 is shifted parallel to the x and y axes so that its initial point is z_1 , the resulting terminal point is $z_1 + z_2$. If 0, z_1 and z_2 are not collinear this is the so-called parallelogram law; see below.



The geometric method for obtaining the product $z_1 z_2$ is somewhat more complicated. If we form a triangle with two sides given by the vectors (originating from 0 to) 1 and z_1 , and then form a similar triangle with the same orientation and the vector z_2 corresponding to the vector 1, the vector which then corresponds to z_1 will be $z_1 z_2$.

This can be verified geometrically but will be most transparent when we introduce polar coordinates later in this section. For the moment, we

Layout Samples

Sec. 10.4] SEQUENCES 595

Example 4.

Sequence	Limit points at:	Convergent or divergent
$1, 2, 3, \dots$	(none)	divergent
$\frac{1}{2}, \frac{2}{3}, \frac{3}{4}, \dots$	1	convergent
$\frac{1}{2}, 2, \frac{1}{3}, 3, \frac{1}{4}, 4, \dots$	0	divergent
$\frac{1}{4}, \frac{3}{4}, \frac{1}{5}, \frac{4}{5}, \frac{1}{6}, \frac{5}{6}, \dots$	0 and 1	divergent

A number which appears infinitely often in a sequence is to be regarded as a limit point; this is a matter of convenience and convention.

A sequence z_1, z_2, \dots is said to be **bounded**, if there is a positive number K such that all the terms of the sequence lie in a disk of radius K about the origin, that is,

$$|z_n| < K \quad \text{for all } n.$$

For example, the second and the last sequences in Ex. 4 are bounded while the first and third are not. We observe that the two bounded sequences have limit points. This illustrates the following important theorem.

Theorem 2 (Bolzano⁴ and Weierstrass⁵). *A bounded infinite sequence has at least one limit point.*

Proof. It is obvious that both conditions are necessary: a finite sequence cannot have a limit point, and the sequence $1, 2, 3, \dots$, though infinite, has no limit point because it is not bounded. To prove the theorem, consider a bounded infinite sequence z_1, z_2, \dots and let K be such that $|z_n| < K$ for all n . If only finitely many values of the z_n are different, then, since the sequence is infinite, some number z must occur infinitely many times in the sequence, and, by definition, this number is a limit point of the sequence.

We may now turn to the case when the sequence contains infinitely many different terms. We draw the large square Q_0 in Fig. 293 which contains all z_n . We subdivide Q_0 into four congruent squares. Clearly, at least one of these squares (each taken with its




Fig. 292. Last sequence in Example 4.

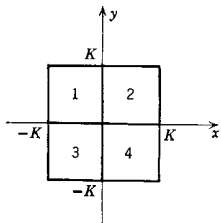


Fig. 293. Proof of Theorem 2.

⁴ BERNHARD BOLZANO (1781–1848), German mathematician, a pioneer in the study of point sets.

⁵ Cf. footnote 3 in Sec. 10.3.



Sec. 10.4] SEQUENCES 595

Example 4.

Sequence	Limit points at:	Convergent or divergent
$1, 2, 3, \dots$	(none)	divergent
$\frac{1}{2}, \frac{2}{3}, \frac{3}{4}, \dots$	1	convergent
$\frac{1}{2}, 2, \frac{1}{3}, 3, \frac{1}{4}, 4, \dots$	0	divergent
$\frac{1}{4}, \frac{3}{4}, \frac{1}{5}, \frac{4}{5}, \frac{1}{6}, \frac{5}{6}, \dots$	0 and 1	divergent

A number which appears infinitely often in a sequence is to be regarded as a limit point; this is a matter of convenience and convention.

A sequence z_1, z_2, \dots is said to be **bounded**, if there is a positive number K such that all the terms of the sequence lie in a disk of radius K about the origin, that is,

$$|z_n| < K \quad \text{for all } n.$$

For example, the second and the last sequences in Ex. 4 are bounded while the first and third are not. We observe that the two bounded sequences have limit points. This illustrates the following important theorem.

Theorem 2 (Bolzano⁴ and Weierstrass⁵). *A bounded infinite sequence has at least one limit point.*

Proof. It is obvious that both conditions are necessary: a finite sequence cannot have a limit point, and the sequence $1, 2, 3, \dots$, though infinite, has no limit point because it is not bounded. To prove the theorem, consider a bounded infinite sequence z_1, z_2, \dots and let K be such that $|z_n| < K$ for all n . If only finitely many values of the z_n are different, then, since the sequence is infinite, some number z must occur infinitely many times in the sequence, and, by definition, this number is a limit point of the sequence.

We may now turn to the case when the sequence contains infinitely many different terms. We draw the large square Q_0 in Fig. 293 which contains all z_n . We subdivide Q_0 into four congruent squares. Clearly, at least one of these squares (each taken with its




Fig. 292. Last sequence in Example 4.

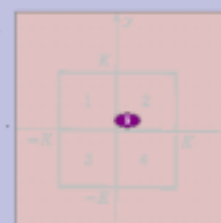


Fig. 293. Proof of Theorem 2.

⁴ BERNHARD BOLZANO (1781–1848), German mathematician, a pioneer in the study of point sets.

⁵ Cf. footnote 3 in Sec. 10.3.

Layout Samples

10.1 Winding Numbers and the Cauchy Residue Theorem

107

PROOF. Since

$$f(z) = \frac{C_{-1}}{z - z_0} + C_0 + C_1(z - z_0) + \dots,$$

$$(z - z_0)f(z) = C_{-1} + C_0(z - z_0) + C_1(z - z_0)^2 + \dots,$$

and

$$\lim_{z \rightarrow z_0} (z - z_0)f(z) = C_{-1}.$$

The second equality in (1) follows since

$$\begin{aligned} \lim_{z \rightarrow z_0} (z - z_0)f(z) &= \lim_{z \rightarrow z_0} (z - z_0) \frac{A(z)}{B(z)} \\ &= \lim_{z \rightarrow z_0} A(z) \frac{B(z) - B(z_0)}{z - z_0} = \frac{A(z_0)}{B'(z_0)}. \quad \square \end{aligned}$$

(ii) If f has a pole of order k at z_0 ,

$$C_{-1} = \frac{1}{(k-1)!} \frac{d^{k-1}}{dz^{k-1}} [(z - z_0)^k f(z)] \text{ evaluated at } z_0.$$

PROOF. Setting

$$f(z) = C_{-k}(z - z_0)^{-k} + \dots + C_{-1}(z - z_0)^{-1} + C_0 + C_1(z - z_0) + \dots$$

$$g(z) = (z - z_0)^k f(z) = C_{-k} + \dots + C_{-1}(z - z_0)^{k-1} + C_0(z - z_0)^k + \dots$$

$$\frac{d^{k-1}g(z)}{dz^{k-1}} = (k-1)!C_{-1} + k!C_0(z - z_0) + \dots$$

and the equality follows. \square

(iii) In most cases of higher-order poles, as with essential singularities, the most convenient way to determine the residue is directly from the Laurent expansion.

EXAMPLES

(i) $\text{Res}(\csc z; 0) = \frac{1}{\cos 0} = 1.$

(ii) $\text{Res}\left(\frac{1}{z^4 - 1}; i\right) = \frac{1}{4i^3} = \frac{i}{4}.$

(iii) $\text{Res}\left(\frac{1}{z^3}; 0\right) = 0.$

(iv) $\text{Res}\left(\sin \frac{1}{z-1}; 1\right) = 1$, since

$$\sin \frac{1}{z-1} = \frac{1}{z-1} - \frac{1}{3!(z-1)^3} + \frac{1}{5!(z-1)^5} - \dots$$



10.1 Winding Numbers and the Cauchy Residue Theorem

107

PROOF. Since

$$f(z) = \frac{C_{-1}}{z - z_0} + C_0 + C_1(z - z_0) + \dots,$$

$$(z - z_0)f(z) = C_{-1} + C_0(z - z_0) + C_1(z - z_0)^2 + \dots,$$

and

$$\lim_{z \rightarrow z_0} (z - z_0)f(z) = C_{-1}.$$

The second equality in (1) follows since

$$\begin{aligned} \lim_{z \rightarrow z_0} (z - z_0)f(z) &= \lim_{z \rightarrow z_0} (z - z_0) \frac{A(z)}{B(z)} \\ &= \lim_{z \rightarrow z_0} A(z) \frac{B(z) - B(z_0)}{z - z_0} = \frac{A(z_0)}{B'(z_0)}. \quad \square \end{aligned}$$

(ii) If f has a pole of order k at z_0 ,

$$C_{-1} = \frac{1}{(k-1)!} \frac{d^{k-1}}{dz^{k-1}} [(z - z_0)^k f(z)] \text{ evaluated at } z_0.$$

PROOF. Setting

$$f(z) = C_{-k}(z - z_0)^{-k} + \dots + C_{-1}(z - z_0)^{-1} + C_0 + C_1(z - z_0) + \dots$$

$$g(z) = (z - z_0)^k f(z) = C_{-k} + \dots + C_{-1}(z - z_0)^{k-1} + C_0(z - z_0)^k + \dots$$

$$\frac{d^{k-1}g(z)}{dz^{k-1}} = (k-1)!C_{-1} + k!C_0(z - z_0) + \dots$$

and the equality follows. \square

(iii) In most cases of higher-order poles, as with essential singularities, the most convenient way to determine the residue is directly from the Laurent expansion.

EXAMPLES

(i) $\text{Res}(\csc z; 0) = \frac{1}{\cos 0} = 1.$

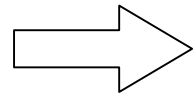
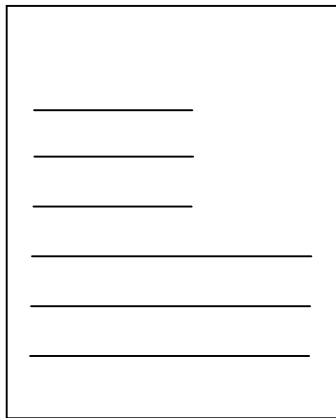
(ii) $\text{Res}\left(\frac{1}{z^4 - 1}; i\right) = \frac{1}{4i^3} = \frac{i}{4}.$

(iii) $\text{Res}\left(\frac{1}{z^3}; 0\right) = 0.$

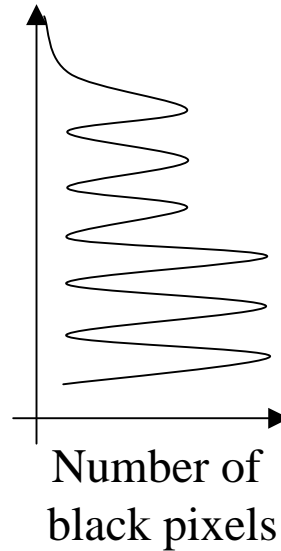
(iv) $\text{Res}\left(\sin \frac{1}{z-1}; 1\right) = 1$, since

$$\sin \frac{1}{z-1} = \frac{1}{z-1} - \frac{1}{3!(z-1)^3} + \frac{1}{5!(z-1)^5} - \dots$$

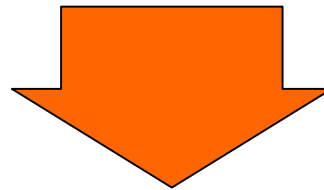
Method of Line Segmentation (1)



Horizontal
projection
histogram



Periodic minimal positions
= Line segment boundary



Math expressions breaks
“periodicity”

Line Segmentation (sample)

Let M be an n -dimensional closed hypersurface in a unit sphere $S^{n+1}(1)$ of dimension $n + 1$. Let S denote the squared norm of the second fundamental form of M . It is well-known that Chern, do Carmo and Kobayashi [2] and Lawson [3] obtained independently that Clifford tori are the only closed minimal hypersurfaces of the unit sphere with $S = n$. When the scalar curvature of M is constant, there are very nice results on the rigidity of the Clifford torus (see [5] and [6]). On the other hand, Otsuki[4] studied the converse problem for minimal hypersurfaces in $S^{n+1}(1)$. He proved that if M is a closed minimal hypersurface in $S^{n+1}(1)$ with two distinct principal curvatures and the multiplicities of them are at least two, then M is $S^m(\sqrt{m/n}) \times S^{n-m}(\sqrt{(n-m)/n})$ ($1 < m < n - 1$). But for the case in which one of the two principal curvatures is simple, he constructed infinitely many minimal hypersurfaces other than $S^1(\sqrt{1/n}) \times S^{n-1}(\sqrt{(n-1)/n})$ which are not congruent to each other in $S^{n+1}(1)$. When professor K. Shiohama visited China in 1993, he proposed the following interesting problem:

PROBLEM. *Let M be a closed minimal hypersurface in $S^{n+1}(1)$ with two distinct principal curvatures λ_1 and λ_2 and one of them be simple (we assume λ_1). Is there a constant $\epsilon = \epsilon(n)$ such that if $|\lambda_1 - \lambda_{10}| < \epsilon$ and $|\lambda_2 - \lambda_{20}| < \epsilon/(n-1)$ then M is $S^1(\sqrt{1/n}) \times S^{n-1}(\sqrt{(n-1)/n})$, where λ_{i0} are the corresponding principal curvatures of $S^1(\sqrt{1/n}) \times S^{n-1}(\sqrt{(n-1)/n})$.*

Line Segmentation (sample)

Let M be an n -dimensional closed hypersurface in a unit sphere $S^{n+1}(1)$ of dimension $n + 1$. Let S denote the squared norm of the second fundamental form of M . It is well-known that Chern, do Carmo and Kobayashi [2] and Lawson [3] obtained independently that Clifford tori are the only closed minimal hypersurfaces of the unit sphere with $S = n$. When the scalar curvature of M is constant, there are very nice results on the rigidity of the Clifford torus (see [5] and [6]). On the other hand, Otsuki[4] studied the converse problem for minimal hypersurfaces in $S^{n+1}(1)$. He proved that if M is a closed minimal hypersurface in $S^{n+1}(1)$ with two distinct principal curvatures and the multiplicities of them are at least two, then M is $S^m(\sqrt{m/n}) \times S^{n-m}(\sqrt{(n-m)/n})$ ($1 < m < n - 1$). But for the case in which one of the two principal curvatures is simple, he constructed infinitely many minimal hypersurfaces other than $S^1(\sqrt{1/n}) \times S^{n-1}(\sqrt{(n-1)/n})$ which are not congruent to each other in $S^{n+1}(1)$. When professor K. Shiohama visited China in 1993, he

$$\begin{aligned}
 & (\sqrt{1/n}) \times S^{n-1}(\sqrt{(n-1)/n}) \\
 & S^1(\sqrt{1/n}) \times S^{n-1}(\sqrt{(n-1)/n})
 \end{aligned}$$

distinct
there a
M is
atures

Line Segmentation (sample)

$$A\sigma(Y) = Y.$$

Since M is étale, Y is contained in $GL_r(O_{\mathcal{E}}^{\dagger})$ by Proposition 3.1.5. Therefore, the assertion of Proposition 5.2.1 follows Lemma 3.3.2.

6. Proof of Lemma 5.2.4. In this section we prove Lemma 5.2.4 using p -adic analysis. Assume that the residue class field k of F is algebraically closed throughout this section.

(6.1) Put $\mathfrak{m}^{\geq \alpha} = \{x \in O_{\widehat{K}alg} \mid |x| \leq |p|^{\alpha}\}$ for any positive number α . Let $O_{\widehat{K}alg}[[z]]$ be the ring of formal power series and denote by ϕ the endomorphism on $O_{\widehat{K}alg}[[z]]$ which is defined by the identity on $O_{\widehat{K}alg}$ and by $\phi(z) = z^p$. We sometimes use the same notation $\mathfrak{m}^{\geq \alpha}$ for the ideal $\mathfrak{m}^{\geq \alpha} O_{\widehat{K}alg}[[z]]$ in $O_{\widehat{K}alg}[[z]]$ for any positive number α . Put $\delta_z = z \frac{d}{dz}$.

Note that, if $V = 1_r + \sum_{n=1}^{\infty} V_n z^n$ (resp. $W = 1_r + \sum_{n=1}^{\infty} W_n z^n$) is a matrix in $GL_r(O_{\widehat{K}alg}[[z]])$ such that $|V_n| \leq \min\{\xi^n, c\}$ (resp. $|W_n| \leq \min\{\xi^n, c\}$) for all $n > 0$ for real numbers $0 < \xi < 1$ and $0 < c < 1$, then VW and V^{-1} also satisfy same conditions.

Line Segmentation (sample)

$$A\sigma(Y) = Y.$$

Since M is étale, Y is contained in $GL_r(O_{\mathcal{E}}^{\dagger})$ by Proposition 3.1.5. Therefore, the assertion of Proposition 5.2.1 follows Lemma 3.3.2.

6. Proof of Lemma 5.2.4. In this section we prove Lemma 5.2.4 using p -adic analysis. Assume that the residue class field k of F is algebraically closed throughout this section.

(6.1) Put $m^{\geq \alpha} = \{x \in O_{\widehat{K}alg} \mid |x| \leq |\alpha|\}$. Let $O_{\widehat{K}alg}[[z]]$ be the ring of formal power series on $O_{\widehat{K}alg}[[z]]$ which is defined by the identity $t = z^p$. We sometimes use the same notation $m^{\geq \alpha}$ for $m^{\geq \alpha} O_{\widehat{K}alg}$ for any positive number α . Put $\delta_z = z \frac{d}{dz}$.

Note that, if $V = 1_r + \sum_{n=1}^{\infty} V_n z^n$ (resp. $W = 1_r + \sum_{n=1}^{\infty} W_n z^n$) is a matrix in $GL_r(O_{\widehat{K}alg}[[z]])$ such that $|V_n| \leq \min\{\xi^n, c\}$ (resp. $|W_n| \leq \min\{\xi^n, c\}$) for all $n > 0$ for real numbers $0 < \xi < 1$ and $0 < c < 1$, then VW and V^{-1} also satisfy same conditions.

Line Segmentation (Sample)

so for r large enough, $J(z) \leq C_2 r^e$ for
 $\lim_{r \rightarrow \infty} J(rz)/r^e$ and $j^*(z) = \lim_{z' \rightarrow z} j(z')$, which
and positively homogeneous of order

(S, \cdot) for resource k and schedule S . If re
leted, we get the corresponding time-const
n $PS \infty |temp, \bar{d}| \sum \sum c_k^v \varphi_{kt} + c_k^f \Delta^+ \varphi_{kt}$. An
oblem is again called *time-optimal*.

Line Segmentation (Sample)

so for r large enough, $J(z) \leq C_2 r^e$ for
 $\lim_{r \rightarrow \infty} J(rz)/r^e$ and $j^*(z) = \lim_{z' \rightarrow z} j(z')$, which
 and positively homogeneous of order

(S, \cdot) for resource k and s
 leted, we get the correspo
 n $PS \infty | temp, \bar{d} | \sum \sum c_k^v \varphi_{kt}$
 oblem is again called *time-*

ding time-
 $+ c_k^f \Delta^+ \varphi_{kt}.$

Method of Line Segmentation (2)

21. Prove that (2) is equivalent to the pair of relations

$$\lim_{z \rightarrow z_0} \operatorname{Re} f(z) = \operatorname{Re} l, \quad \lim_{z \rightarrow z_0} \operatorname{Im} f(z) = \operatorname{Im} l.$$

22. The function $f(z) = 3(z^2 - 1)/(z - 1)$ is not defined for $z = 1$, but for all other values of z it is equal to $3(z + 1)$. Using the definition of the limit, show that $\lim_{z \rightarrow 1} f(z) = 6$. (Note that the limit is established when some formula is found for δ as a function of ϵ .)

21. Prove that (2) is equivalent to the pair of relations

$$\lim_{z \rightarrow z_0} \operatorname{Re} f(z) = \operatorname{Re} l, \quad \lim_{z \rightarrow z_0} \operatorname{Im} f(z) = \operatorname{Im} l.$$

22. The function $f(z) = 3(z^2 - 1)/(z - 1)$ is not defined for $z = 1$, but for all other values of z it is equal to $3(z + 1)$. Using the definition of the limit, show that $\lim_{z \rightarrow 1} f(z) = 6$. (Note that the limit is established when some formula is found for δ as a function of ϵ .)

Section 3

Character/Symbol Recognition

Character/Symbol Recognition

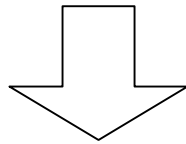
- Character/Symbol recognition
 - Character Recognition in Usual Text Areas
 - Character/Symbol recognition in Math. Expression Areas
- Segmentation of Text / Math. Areas

Character/Symbol Recognition

- Recognition and Segmentation
= Simultaneous Processing
- Use of different OCR engines possibility
 - For text area:
InftyOCR + Toshiba Express Reader (option)
 - For math. Area :
InftyOCR developed in Suzuki Labo.

Character/Symbol Recognition Flow

Input: This value x^2 equals to $y^2 + z^2$

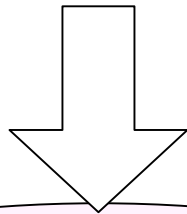


Segmentation by vertical lines

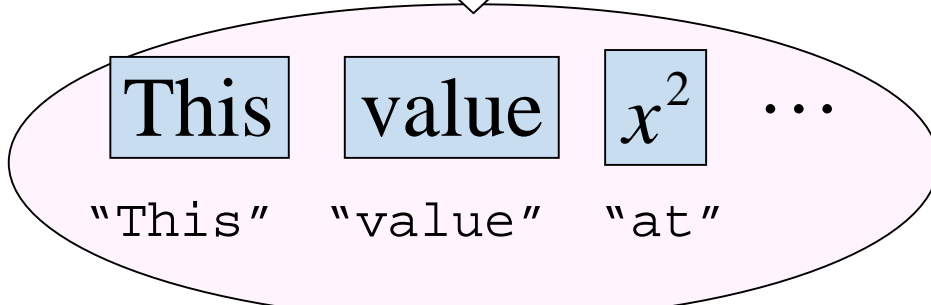
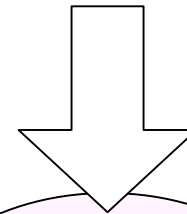
Recognition as "Text characters"

Dictionary

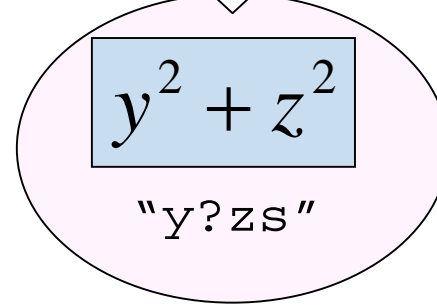
String check



String

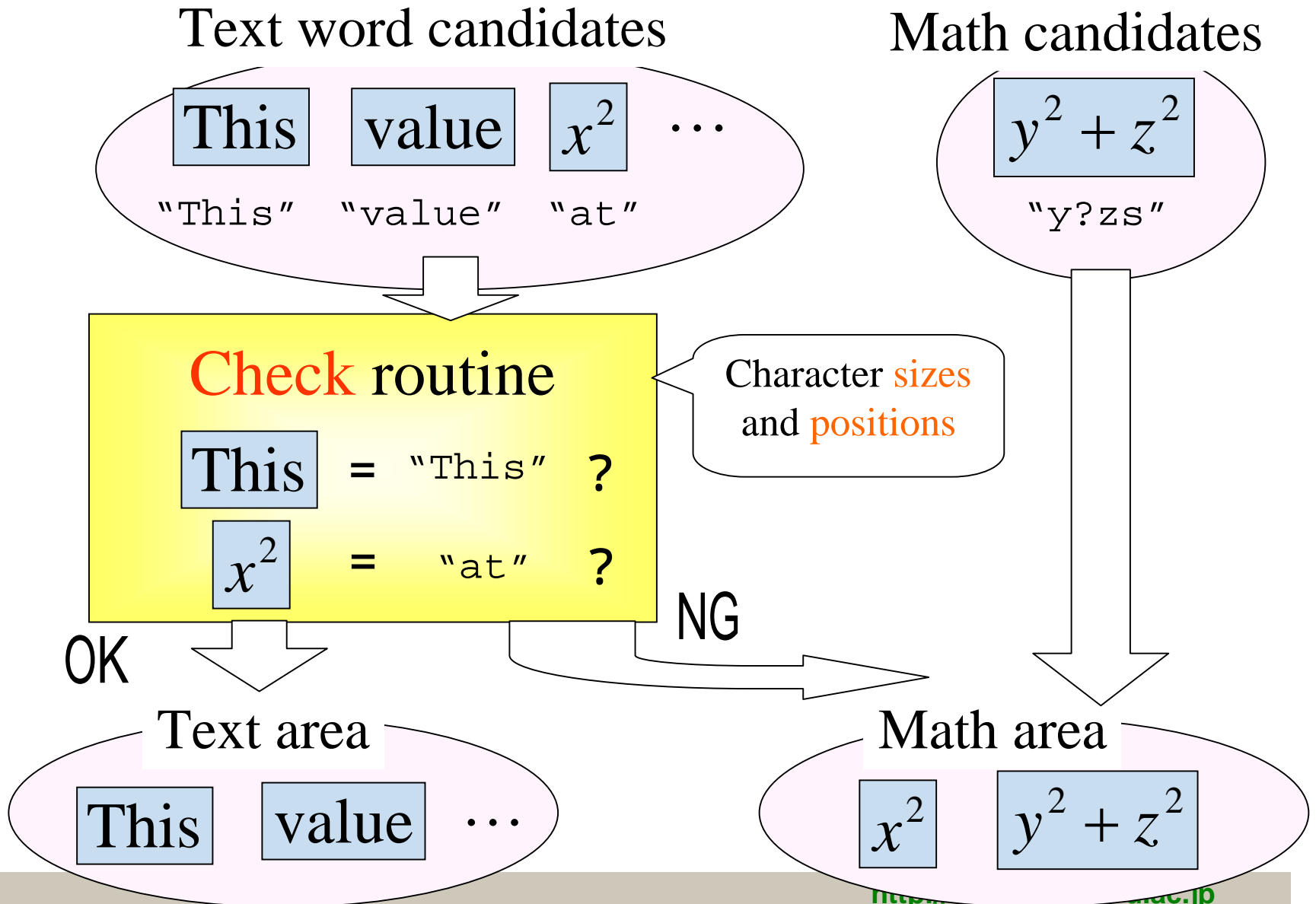


Text (Candidates)

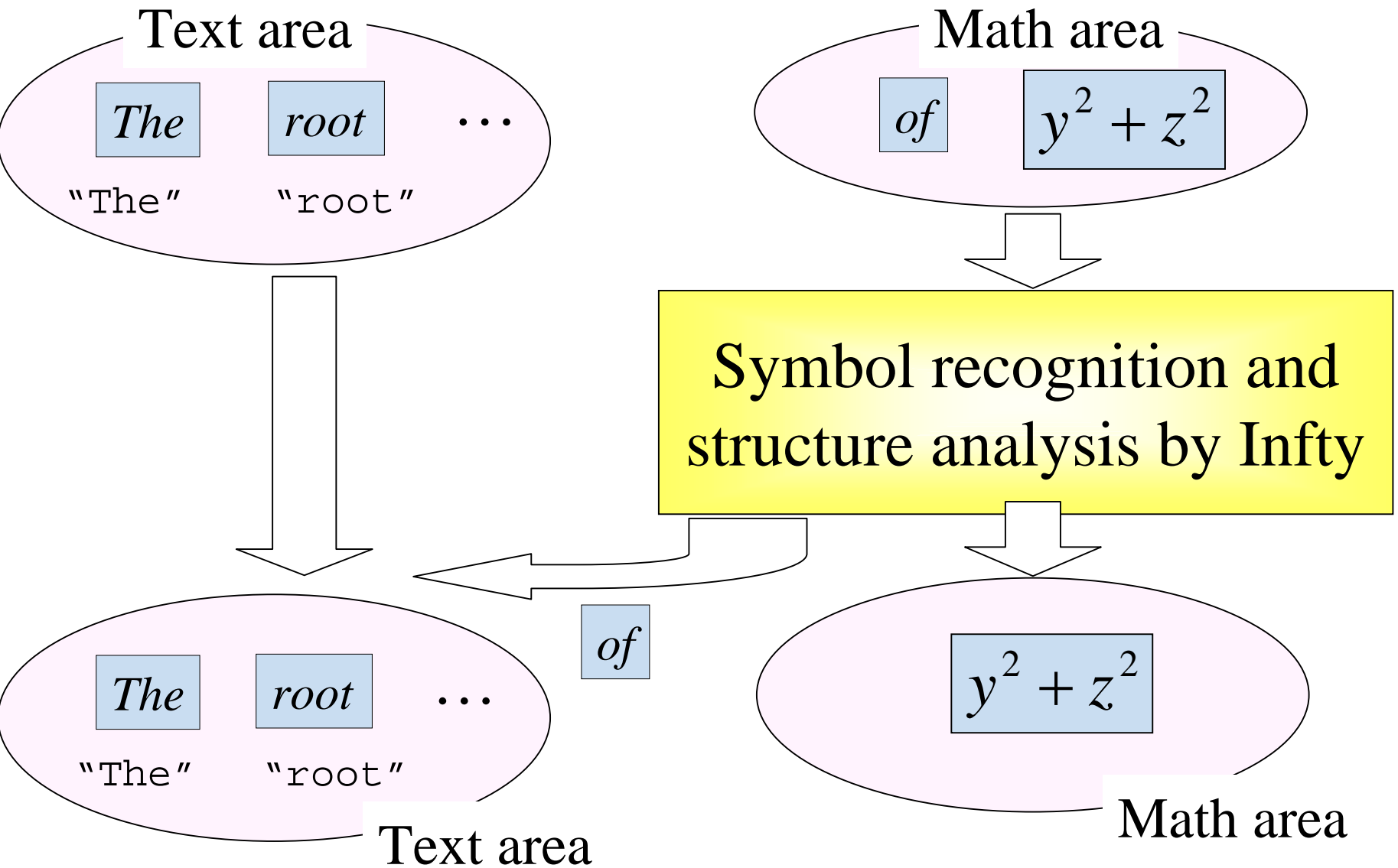


Math (Candidates)

Character/Symbol Recognition Flow



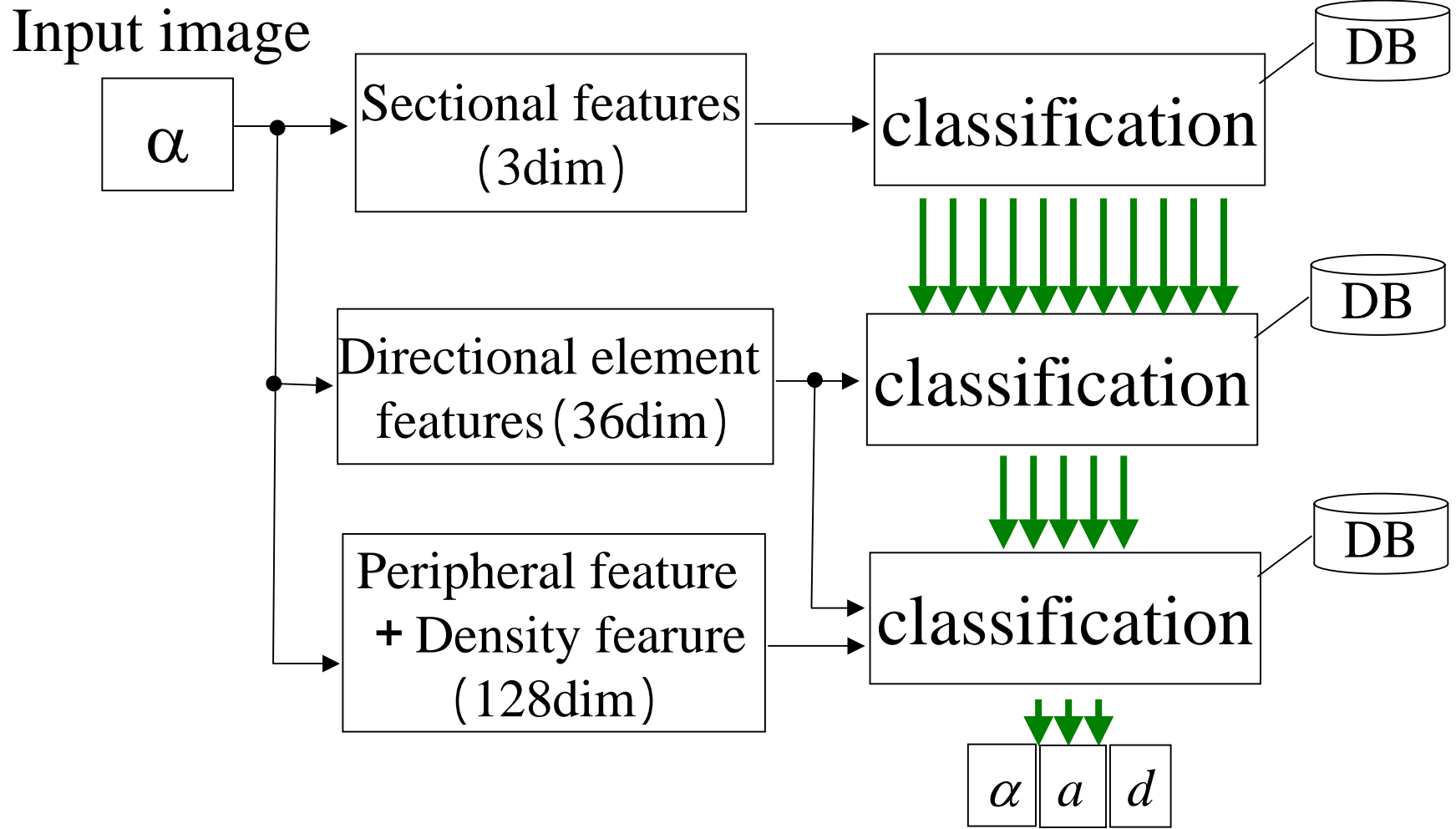
Character/Symbol Recognition Flow



Infty OCR engine

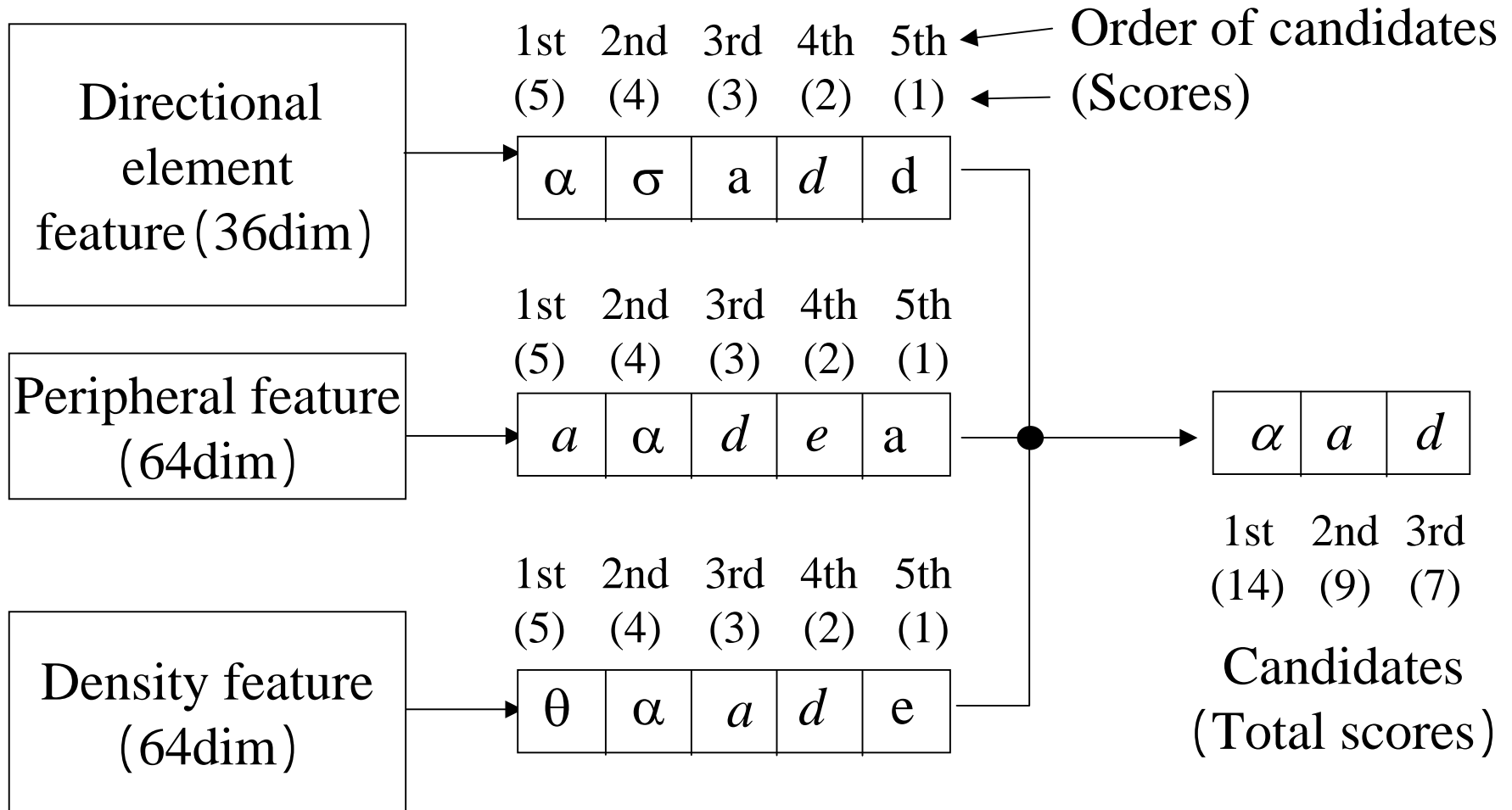
- Developed in Suzuki Lab., using more than 1,500,000 sample images of characters and symbols from various math. books/journals.
- Recognizes more than 500 categories
 - Various math symbols
 - Various fonts: Roman , Italic , Calligraphic , Bbb, some German fonts, etc.
- High speed
 - Three step classification :
“rough” classification “strict” classification

3 step classifications



Recognition result (candidates)

Voting method



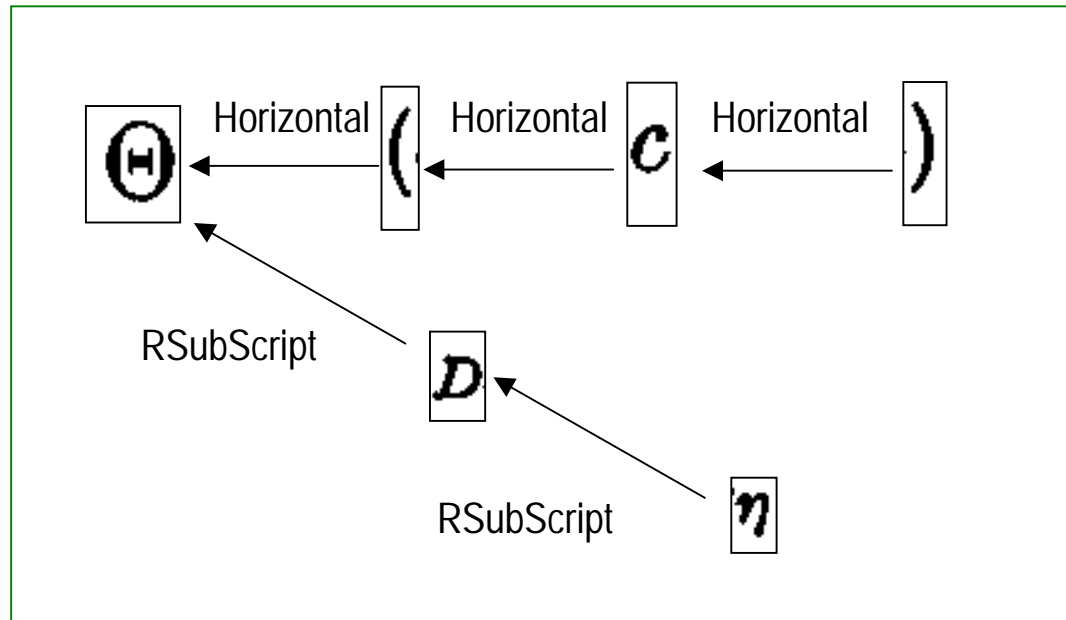
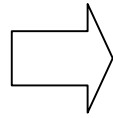
Section 4
Structure Analysis
of formulae

Structure Analysis

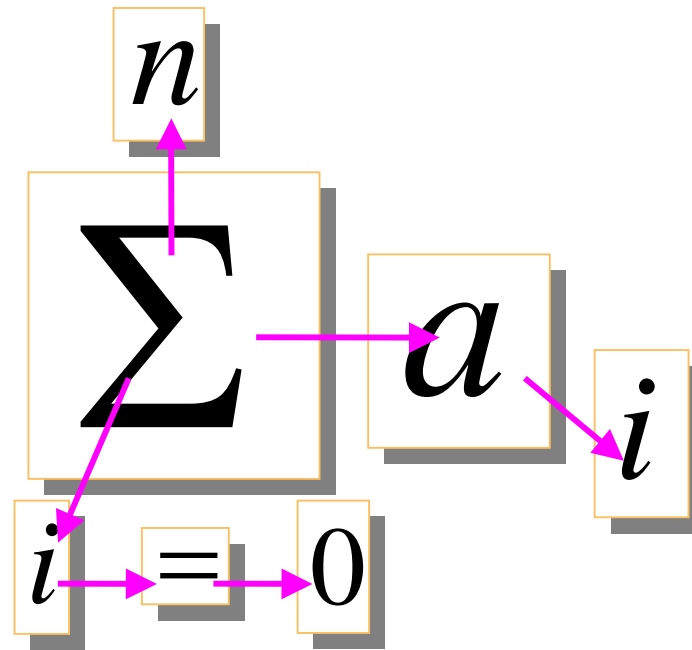
Output (Tree Structure)

Input (image)

$\textcircled{H}_{D\eta}(c)$



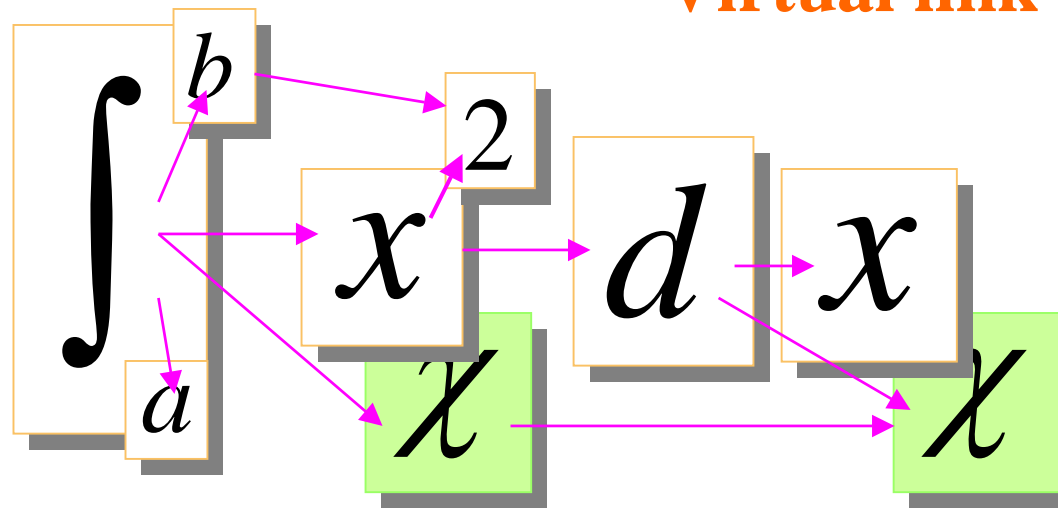
Structure Analysis



Structure Analysis

$$\int_a^b x^2 dx$$

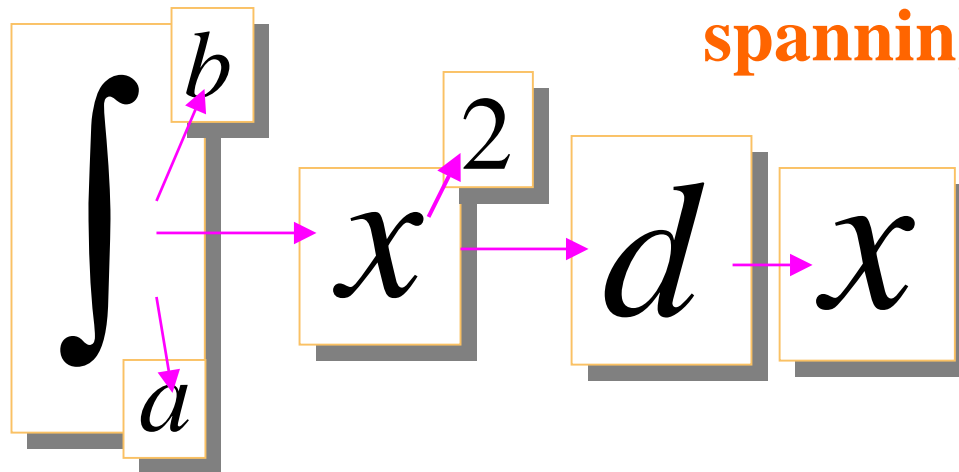
Virtual link network



Structure Analysis

$$\int_a^b x^2 dx$$

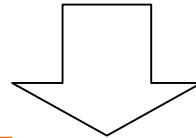
Search for correct
spanning tree



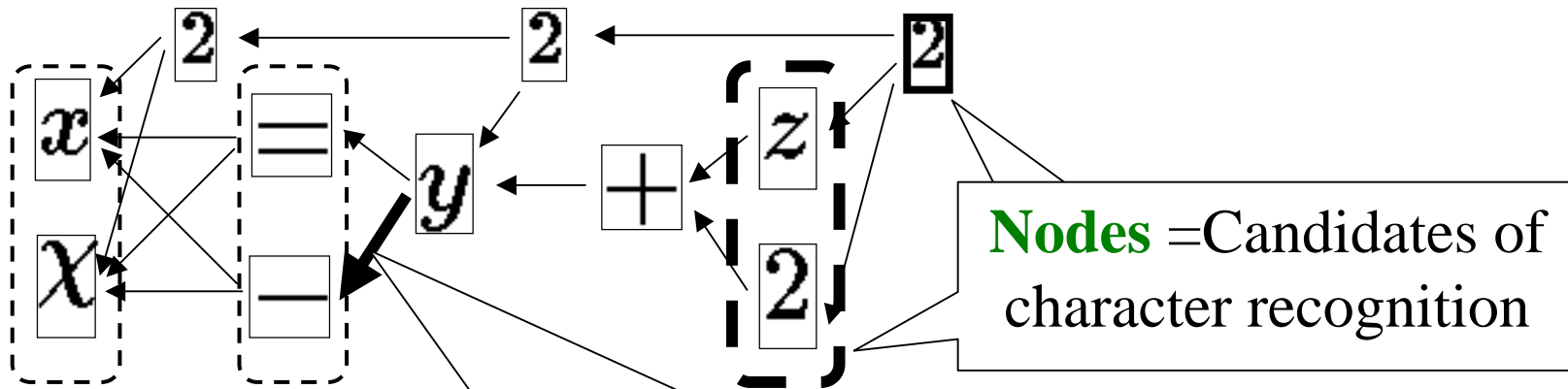
Virtual link network

Input image

$$x^2 = y^2 + z^2$$



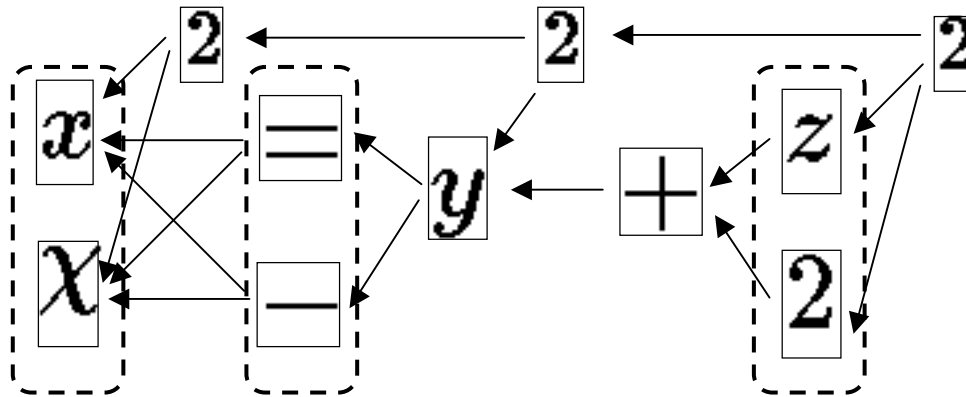
Virtual link network



Each **Link** has a **label** and the link **cost**
Link: Horizontal, Upper, Under, Rsup, Rsub, Lsup, Lsub

Extraction of Structure Tree

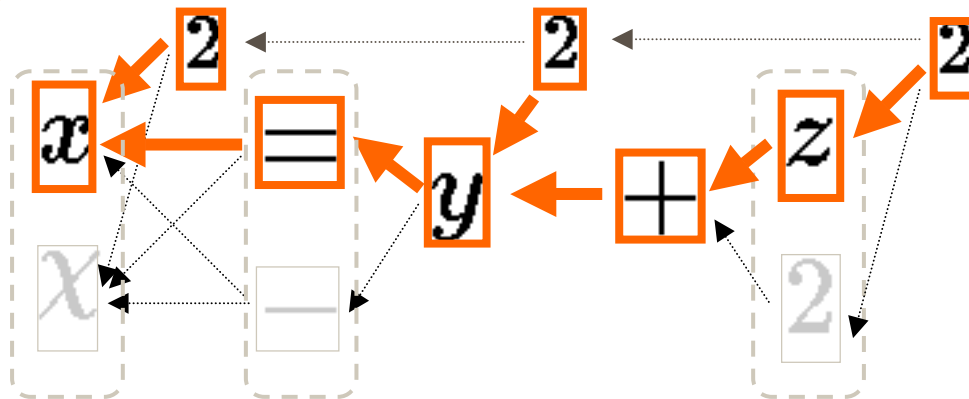
Network



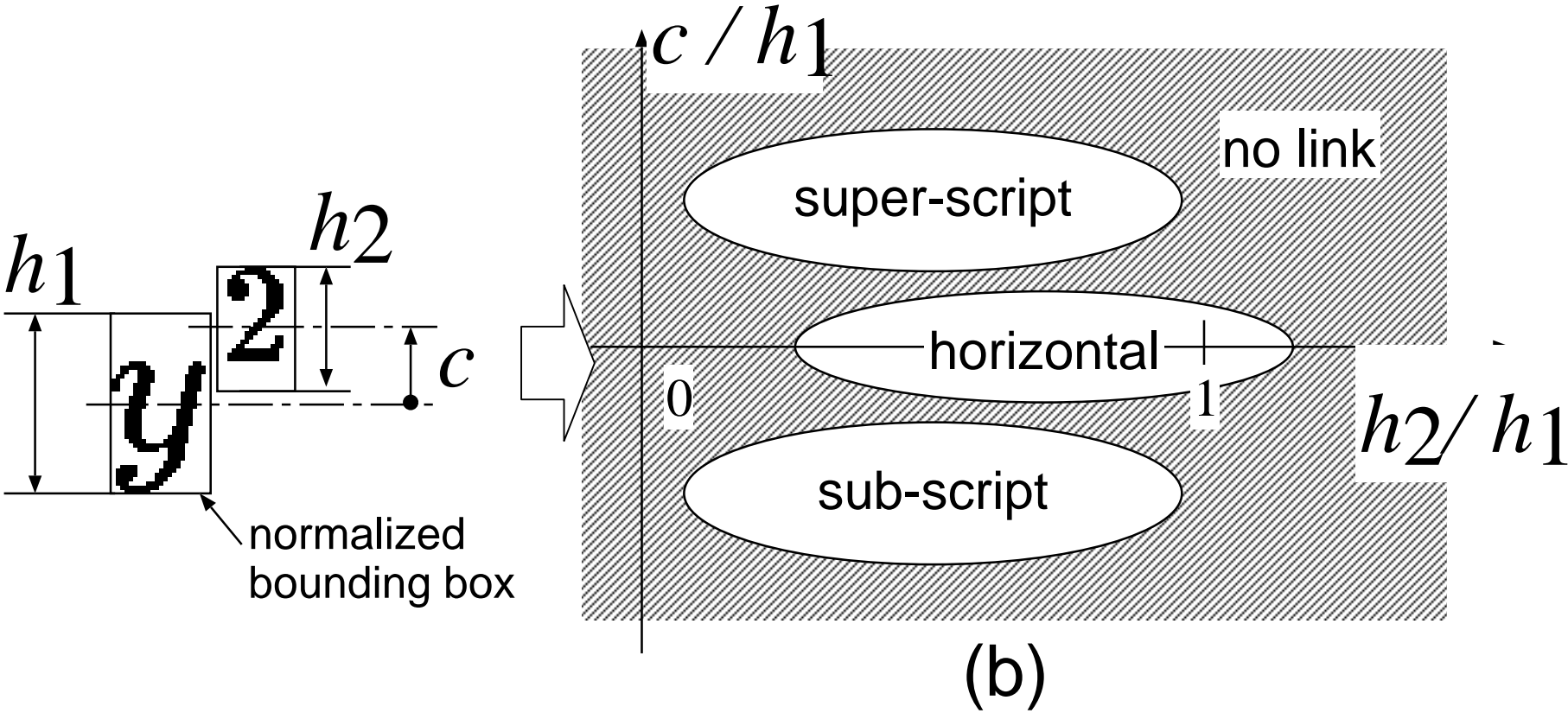
Optimization
under constraints

- Minimum total cost
- Link restrictions

Structure
Tree



Link Cost



Section 5

Ground Truth

(=database with correct recognition results)

Outline of Database

- Extracted from 25 different volumes of mathematics/physics
 - 24 mathematical articles
 - Bulletin of American Math. Soc.,
 - Bulletin de la Soc. Math. France,
 - etc.,
 - 1 chapter of a book of physics
 - Years: 1970 ~ 1999
- Total page number = 485
 - Math articles 453 pages + Physics book 32 pages
- Total character number 700,000
 - Text Area: 550,000 + Math.Area: 150,000

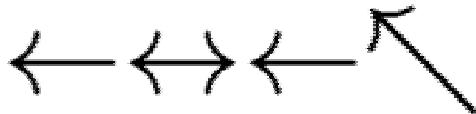
Outline of Database (XML)

- Each character/symbol has:
 - Char code (category , font, id)
 - Math/Text attribute
 - Touched/Broken/Normal attribute
- Tree structure for all math. expressions in DB
 - Link attribute

Symbol category samples (1)

^ ~ _ " \ ˇ

Accent



Arrow

Σ \int Π

Big Symbol

A B C D E F

Blackboard Bold

A B C D E F

Calligraphic

Symbol category samples (2)

A B C a b c

German

$\Gamma \Delta \Theta \alpha \beta \gamma$

Greek

A B C a b c

Italic

012345

Numeral

+ - × / <

Operator

Symbol category samples (3)

#%∞∀∃†

Others

(){}[]

Parenthesis

, · “

Point

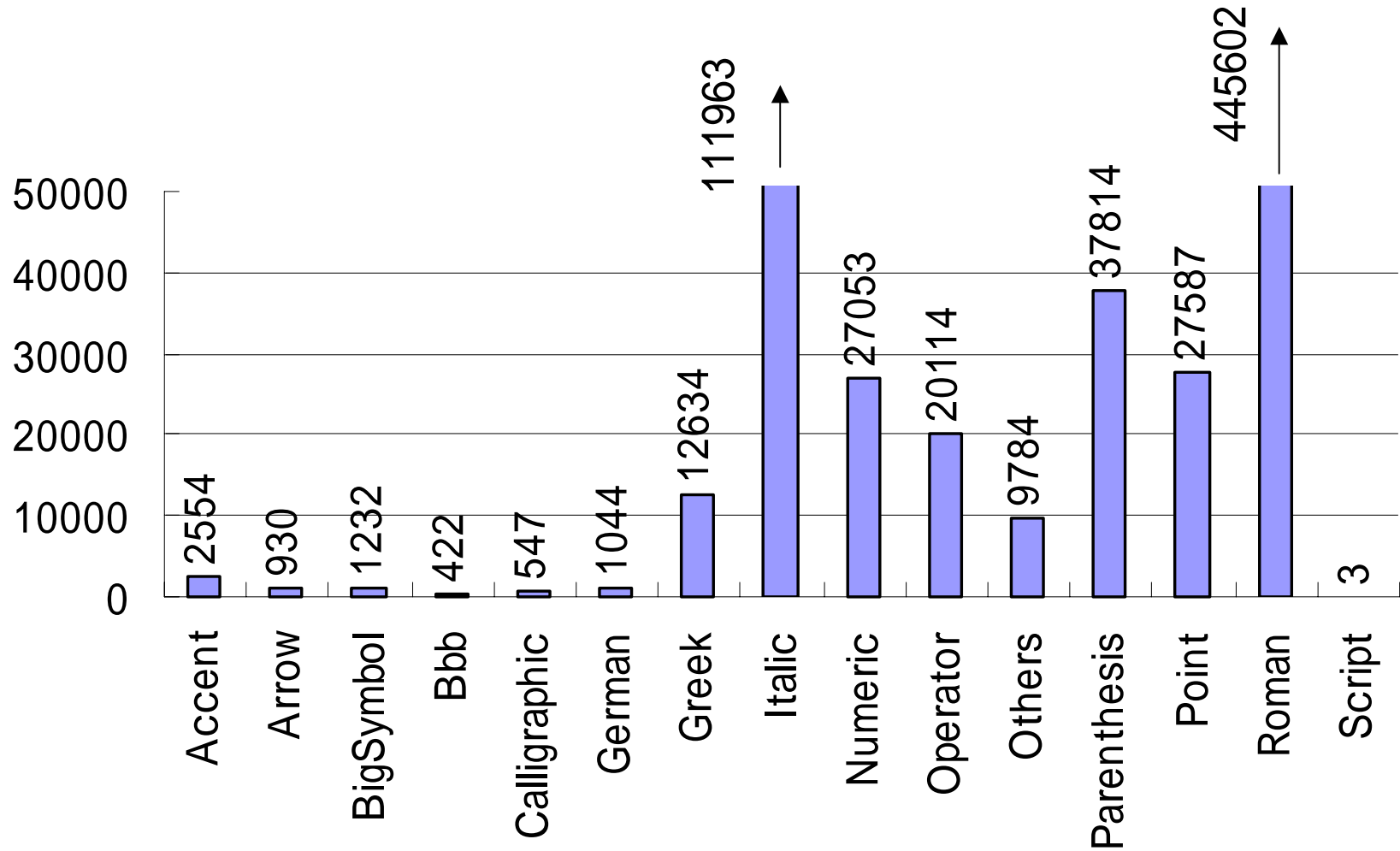
ABCabc

Roman

A B C D E F

Script

Characters in DB per category

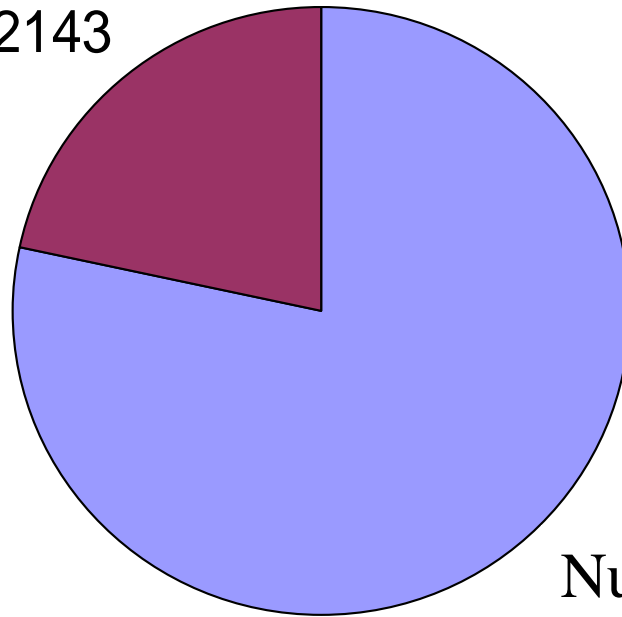


Many italics , parenthesis, ...

Characters in DB in Math/Text areas

Number of characters/symbols
in math. area

152143



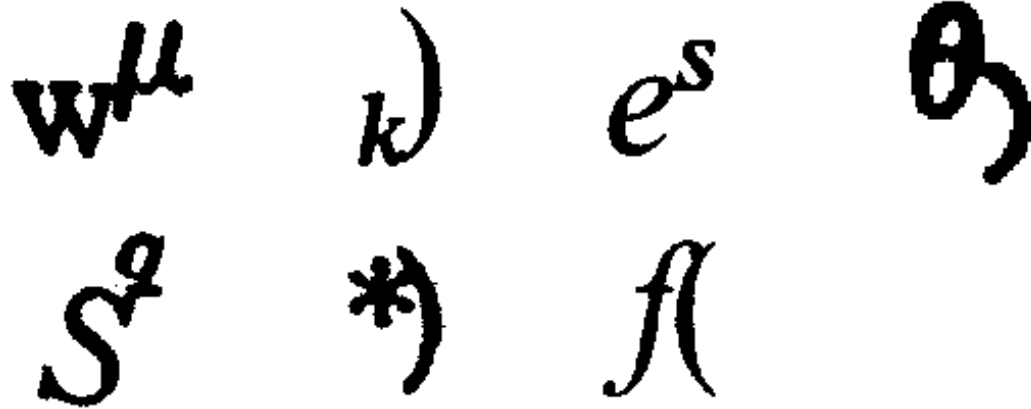
Number of characters
in text area
547140



Text area : Math.area 10 : 3

Abnormal characters

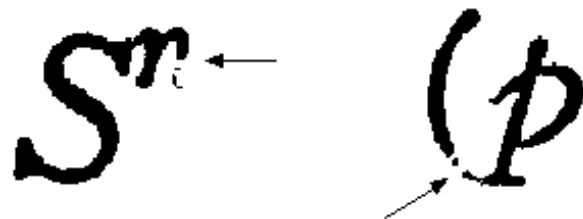
Touched
characters



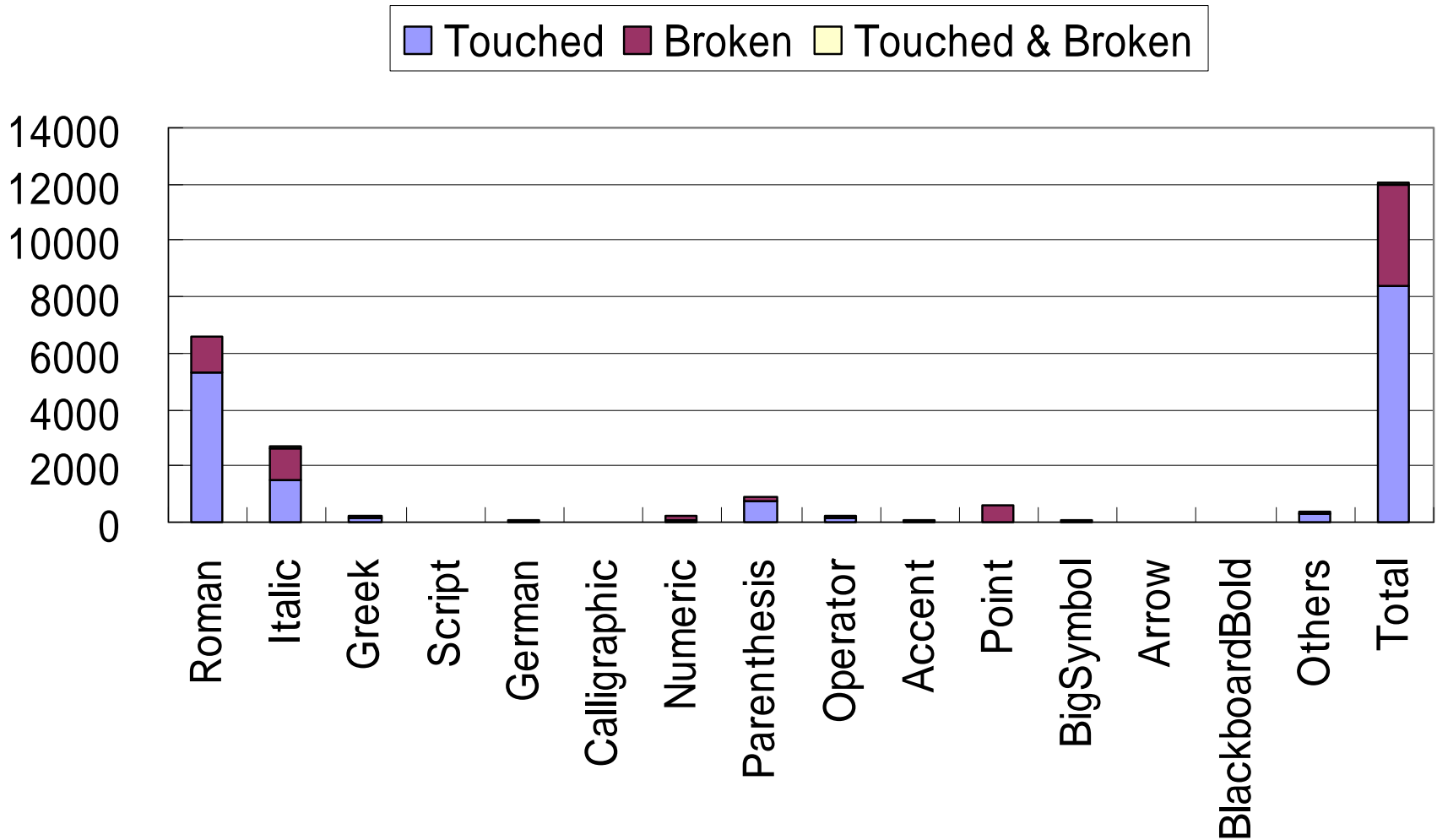
Broken
characters



Touched
and broken

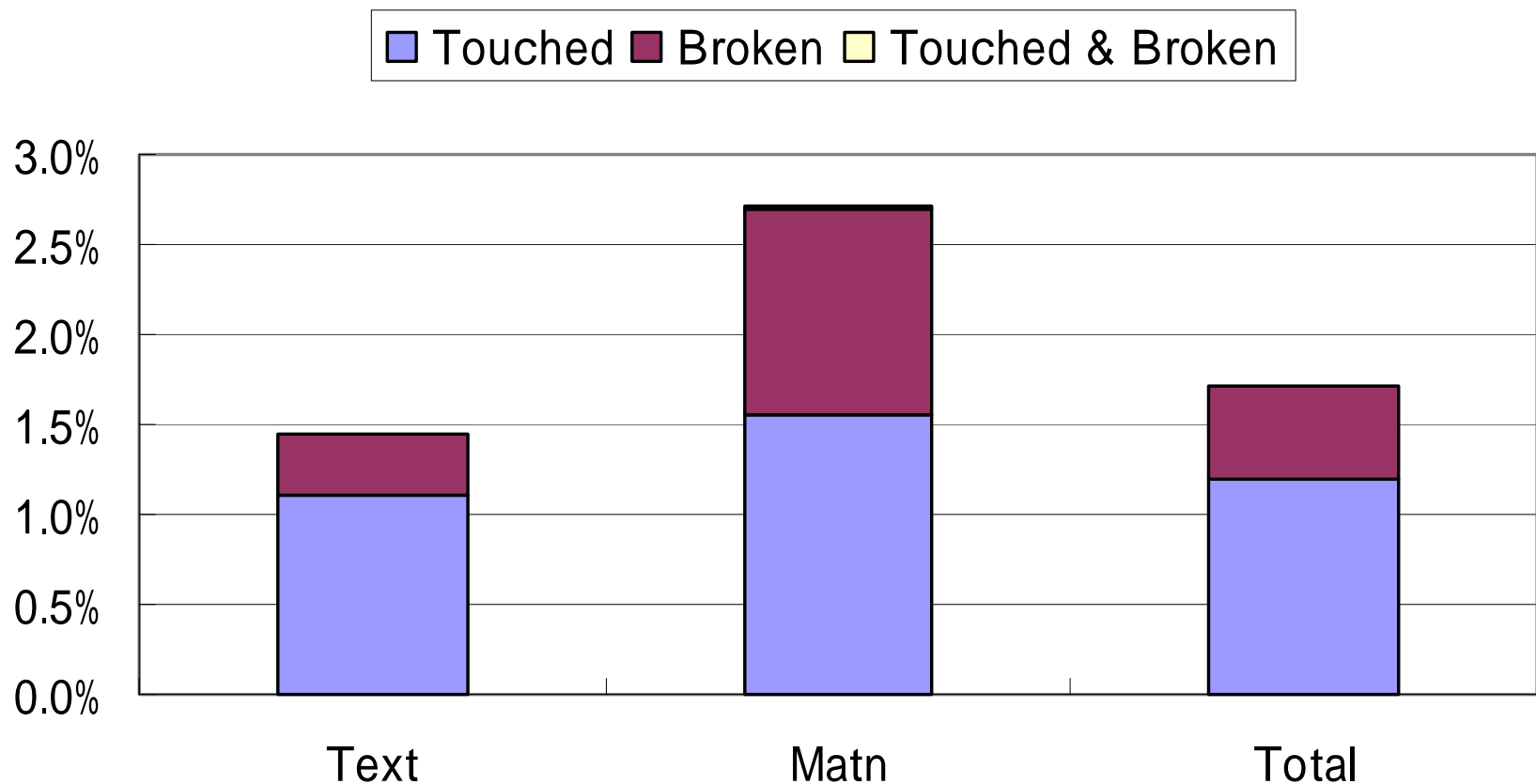


Number of abnormal characters



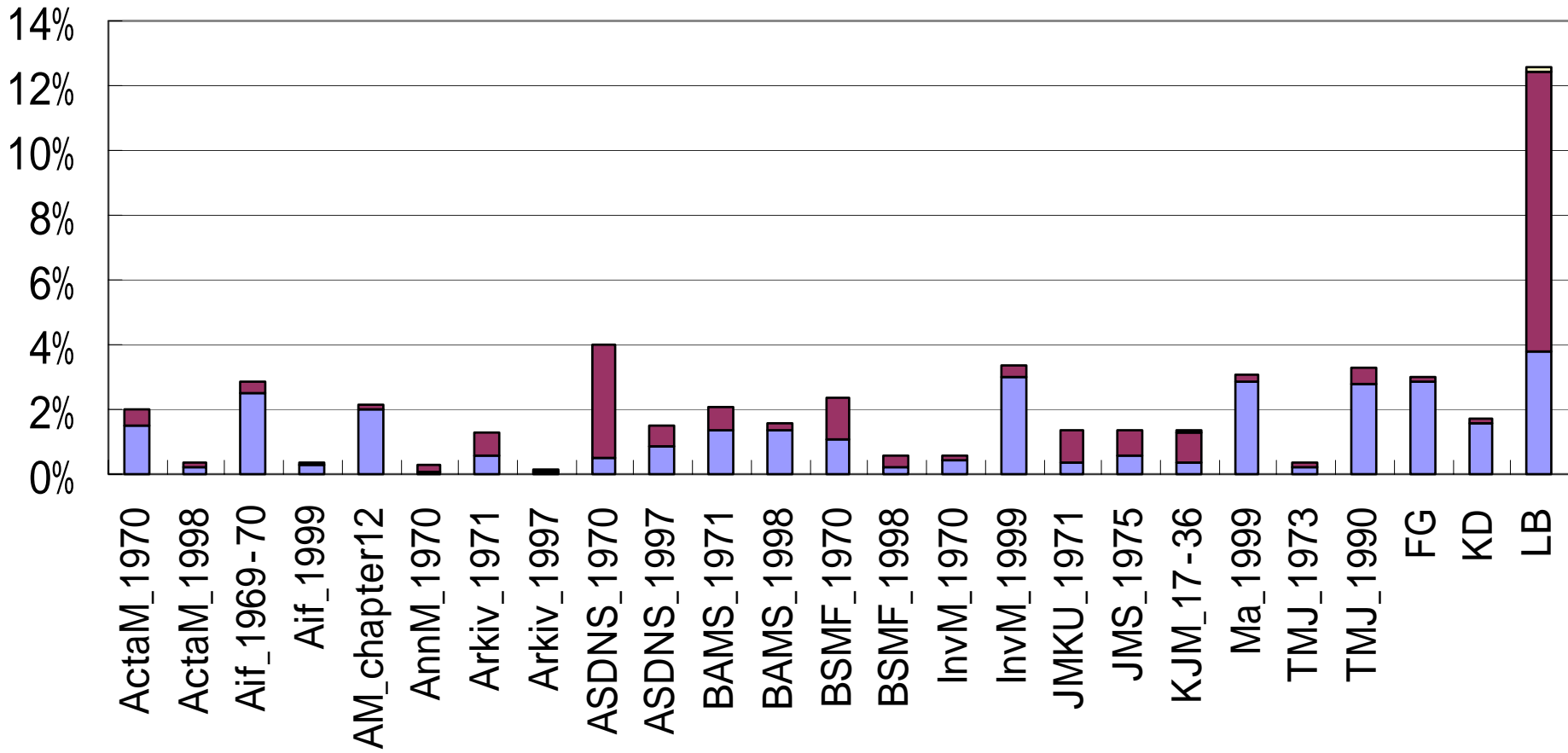
3/4 of abnormal characters are “touched”

Number of abnormal characters



The percentage of abnormal characters in math areas is usually larger than that in text areas.

Number of abnormal characters



Math expressions in the DB

Number of math expressions having 2-dimensional layout structure about 12,300

No math. structure

We assert that X is torsion-free. Indeed, if X is not torsion-free then it has a direct summand $C(p^k)$, $1 \leq k < \infty$ ([2], p. 80), $X = C(p^k) \oplus X'$
This implies that

$$X/p^{k+1}X \cong C(p^k) \oplus X'/p^{k+1}X'$$

which is contrary to

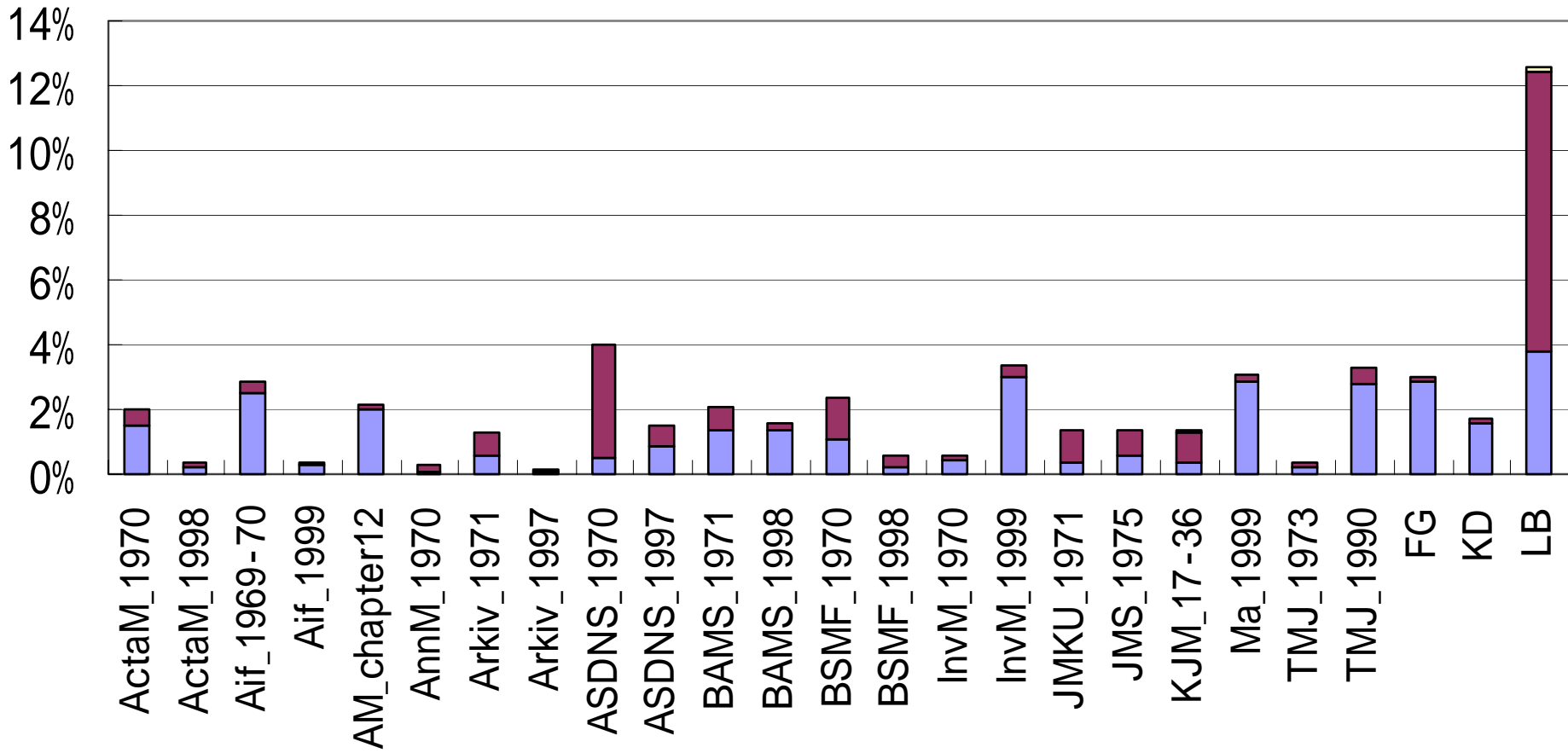
$$X/p^{k+1}X \cong C(p^{k+1})$$

Math. structure

Section 6

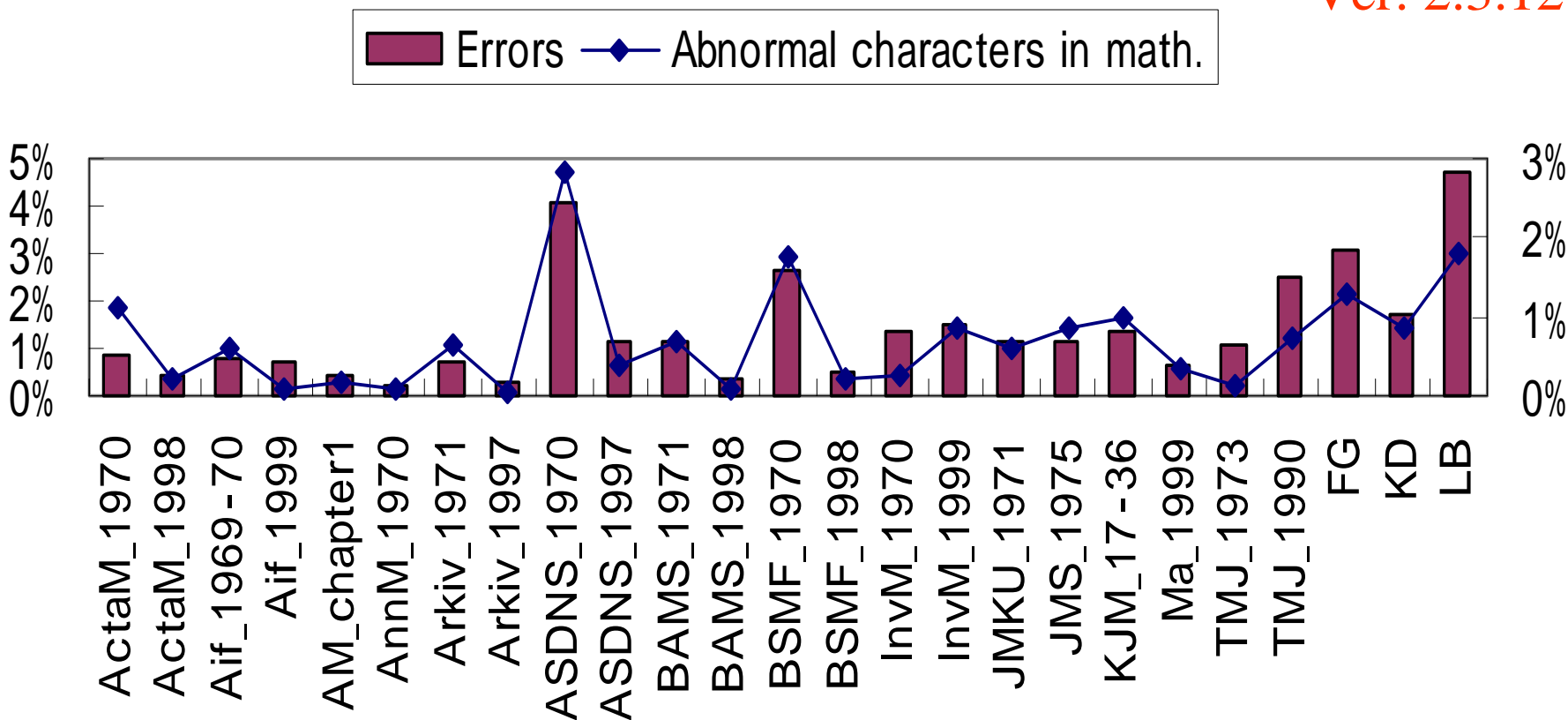
Experimental results

Number of abnormal characters



Error rates and abnormal characters in Math.

Ver. 2.3.12



Ave. recog. rates for 13 papers with abnormal characters < 1% in Math Area :

Text area 99.82% (359030)

Math area 97.42% (65233)

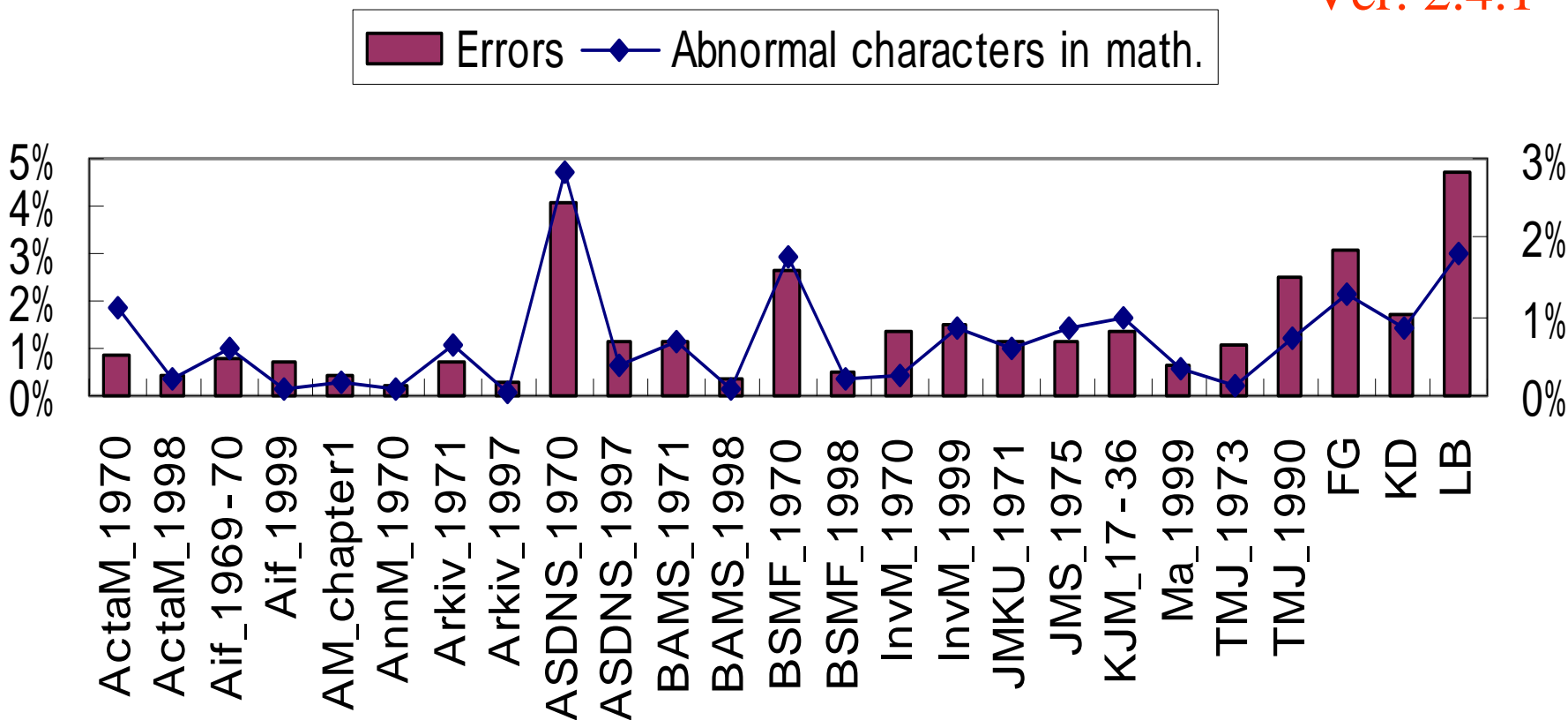
Total 99.39% (424263)

Link error rate in math area 1.07%

Ave. number of chars in math 9

Error rates and abnormal characters in Math.

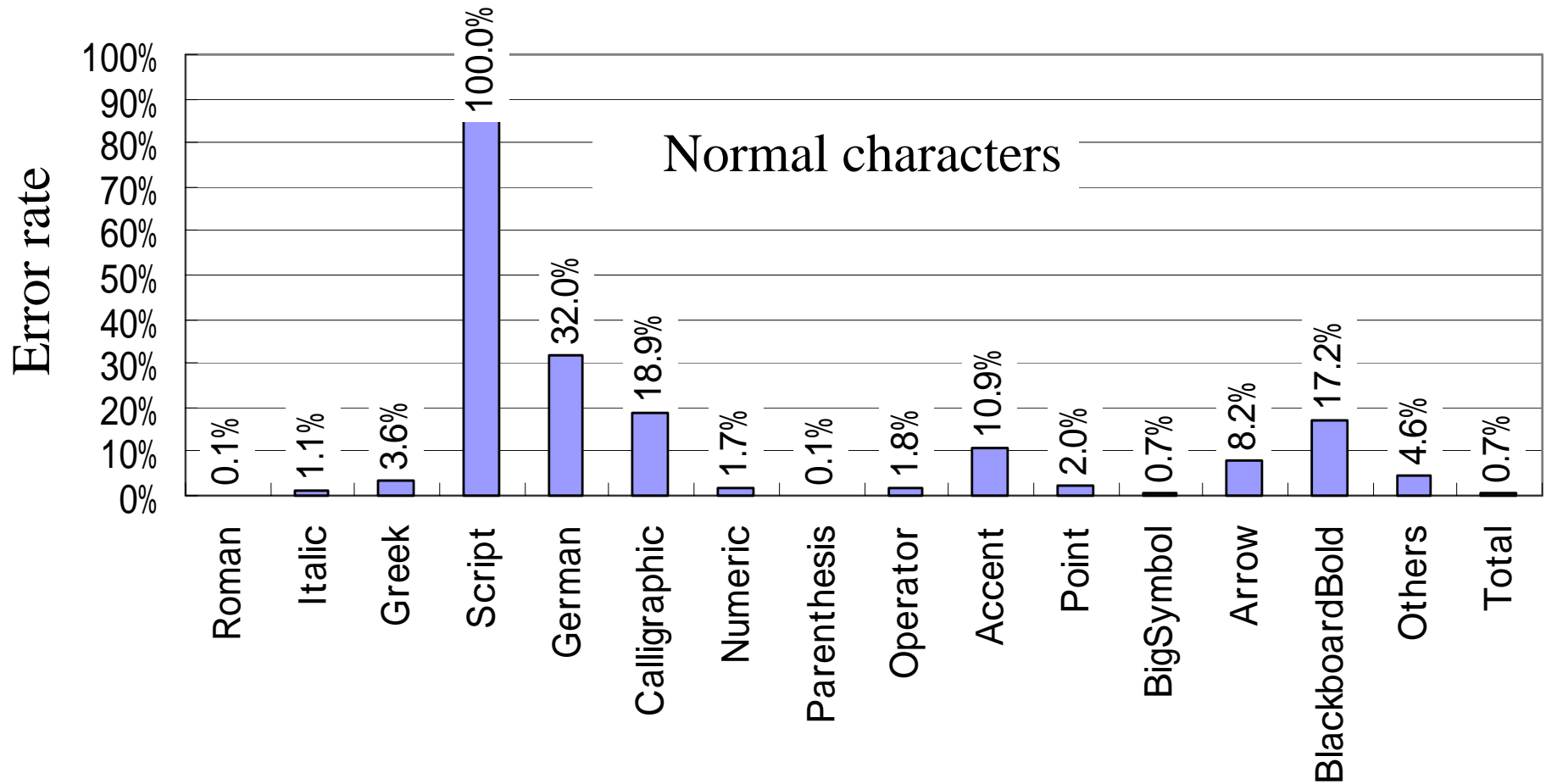
Ver. 2.4.1



Ave. recog. rates for 13 papers with abnormal characters < 1% in Math Area :

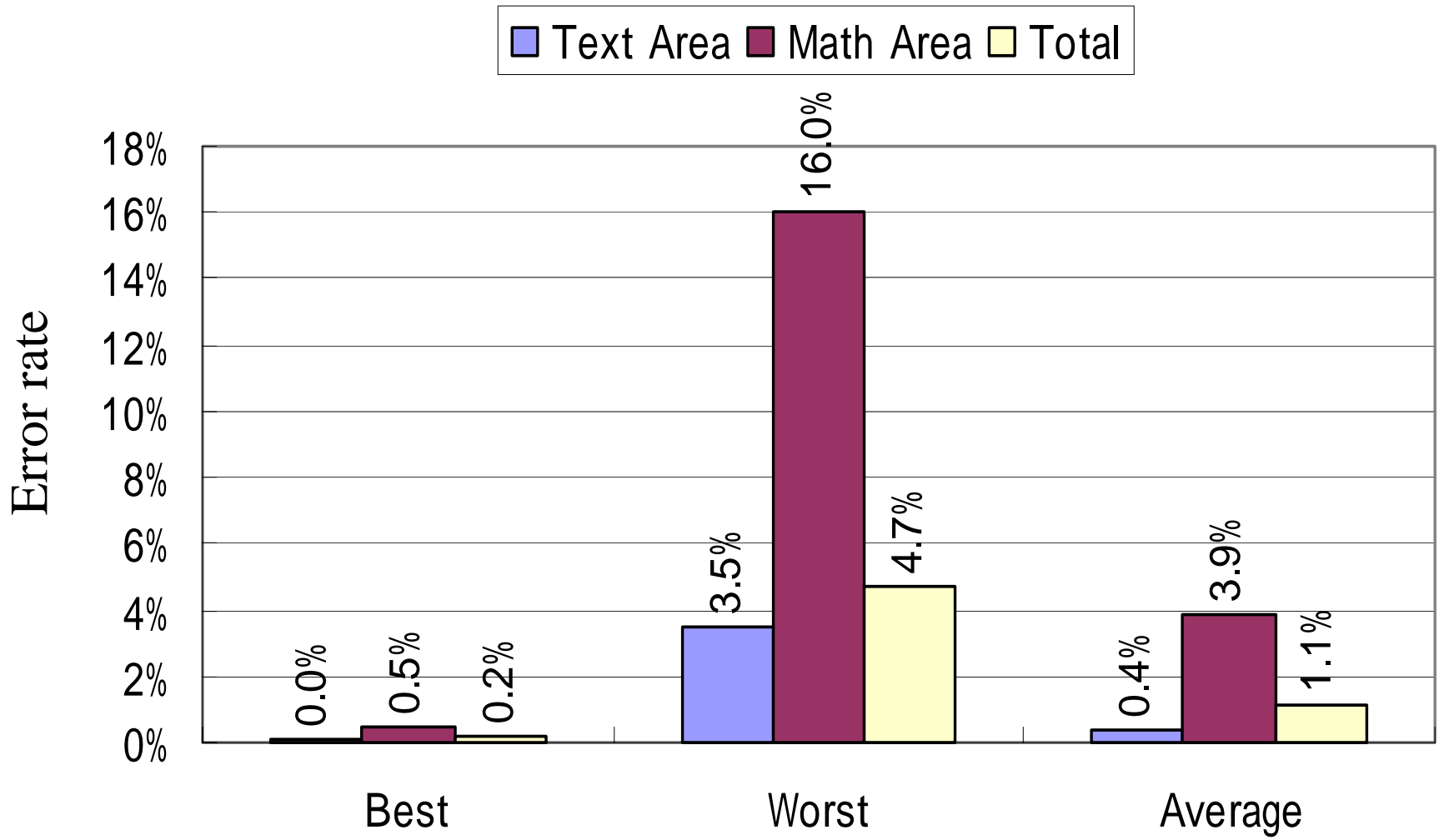
Text area	99.82% (359030)	99.91%
Math area	97.42% (65233)	98.53%
Total	99.39% (424263)	99.58%

Error rates per category



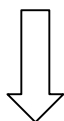
- Error rate of Upright Roman font is 0.1%

Best case and worst case



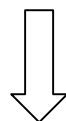
Error samples in math. area

$$\sum_{j=1}^{p-1}$$



$$\sum_{J \approx 1}^{p-1}$$

$$\sum_{p \in \mathfrak{B}_i}$$



$$\sum_{p \in P_i}$$

$$K_t^*$$



$$K_t^*$$

$$I_*^{G_r}$$



$$I_*^{G_Y}$$

(YはUpsilon)

$$\int_{U^*}$$

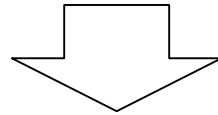


$$\int_U$$

Structure Analysis

Error rate of math. links

(8 papers with abnormal characters < 1%)



1.07 % (=324/30,143)

Success
case 1

$$\frac{(n-1)!}{2\pi^n} \int_{\|a\|=1} \omega_{2n-1} \int_{s_0}^r \frac{\ln|f(sw_0 + \eta sa)| ds}{s} - \int_{s_0}^r \frac{\ln|f(sw_0)| ds}{s}$$

Success
case 2

$$\begin{aligned} \omega &= \frac{\beta}{(-\rho)} + \frac{\gamma}{(-\rho)^2} = i \frac{\sum_{j=1}^n e_j \wedge \bar{e}_j}{(-\rho)} + i \frac{\partial \rho \wedge \bar{\partial} \rho}{(-\rho)^2} \\ &= a i e_1 \wedge \bar{e}_1 + b i \sum_{j=2}^n e_j \wedge \bar{e}_j, \end{aligned}$$

Structure Analysis

■ Error Samples:

$$\boxed{\Theta_{D\eta}(c)} \quad \Rightarrow \quad \Theta_{D\eta}(c)$$

$$\boxed{\int_{c_P^{n-1}} \frac{\alpha_0^{n-1}}{(n-1)!}} \quad \Rightarrow \quad \int_{c_P^n-1} \frac{\alpha_0^{n-1}}{(n-1)!}$$

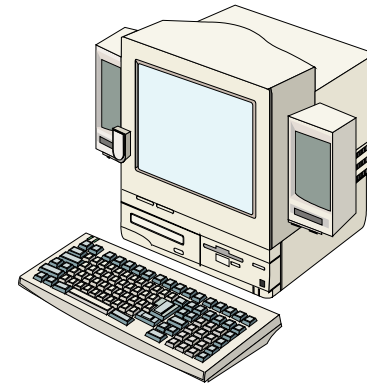
Effect of an
error $C \rightarrow c$

Processing speed

- Process time per page:



About 20 sec.
(Pentium III, 700MHz)



About 10 sec.
(Pentium III, 1GHz)

Section 7

Future problems

Problems

- Improvement of character/symbol recognition for italic fonts and various math symbols.
- Segmentation of touched characters in math areas and unification of broken characters in math areas.
- Analysis of document structure and automatic generation of hyperlinks.
- Extraction of bibliographic data from reference table.
- Etc., Many interesting problems are still unsolved!

Thanks you!

Masakazu Suzuki
Graduate School/ Faculty of Math.
Kyushu University 36, Fukuoka
812-8581 Japan
E-mail: suzuki@math.kyushu-u.ac.jp
<http://math.kyushu-u.ac.jp/suzuki>

<http://infty.math.kyushu-u.ac.jp>

<http://infty.kyushu-u.ac.jp>