Mathematical formulae recognition and logical structure analysis of mathematical papers

DML 2010

July 7, 2010, Paris

Masakazu Suzuki
Kyushu University
InftyProject ((http://www/inftyproject.org)
Science Accessibility Net (http://www.sciaccess.net)

Plan of the talk

- About InftyProject
- Making Rich Digital Mathematical Libraries
  - Process Flow and Technical Components
- Formulae Recognition
- Adaptive Method
  - Character and Symbol Recognition
  - Logical Structure Analysis

Section 1

About Infty Project

InftyProject

- The beginning :
  - Started as a research project to help visually impaired people in scientific fields in 1995.
  - Digitization of of mathematical journals, books, etc..
- Current research subjects :
  - Recognition and understanding of math documents,
  - User interface and data conversion, etc.
- Policy:
  - Priority in practical system development.

InftyProject

- Main system development

  InftyReader : Math OCR software

  InftyEditor : Editor of math documents
  Data conversion（XML, LaTeX, HTML, PDF, etc.)

  ChattyInfty : InftyEditor + speech output

- URL : http;//www.inftyproject.org

Go

sAccessNet

- *InftyReader* is used for
  - Helping people with visual handicaps working in scientific fields,
  - Digitization of mathematical/scientific Journals in Japan,

    e.g. J.Math.Soc.Japan, Japanese J. Math., Tokyo J.Math, etc.,
    (11 journals of mathematics pulished in Japan)

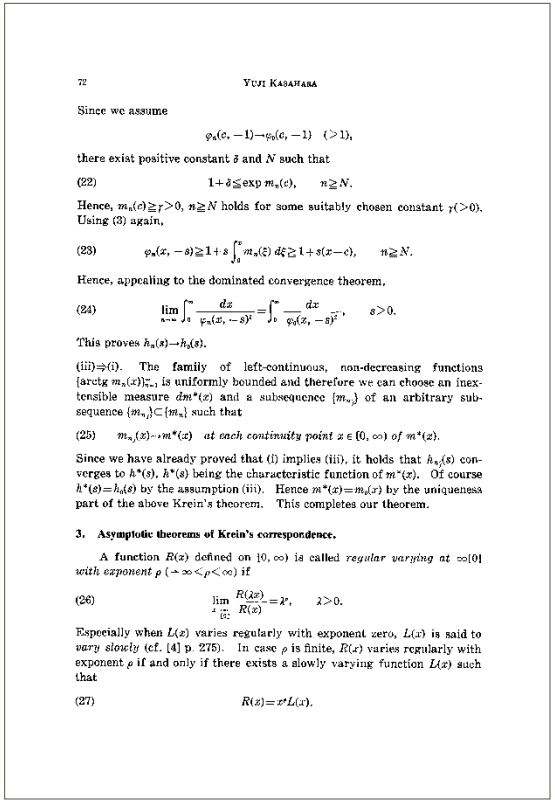
  by the not-for-profit organization “Science Accessibility Net”
- http;//www.sciaccess.net/

Go

“InftyReader” OCR software for math documents

- Demonstration.

Original Image



72 YUJI KASAHARA

Since we assume

$$\varphi_n(c, -1) \to \varphi_0(c, -1) \quad (>1),$$

there exist positive constant $\bar{\delta}$ and N such that

(22) $$1+\bar{\delta} \leqq \exp m_n(c), \qquad n \geqq N.$$

Hence, $m_n(c) \geqq \gamma > 0$, $n \geqq N$ holds for some suitably chosen constant $\gamma(>0)$. Using (3) again,

(23) $$\varphi_n(x, -s) \geqq 1 + s\int_0^x m_n(\xi)\, d\xi \geqq 1 + s(x-c), \qquad n \geqq N.$$

Hence, appealing to the dominated convergence theorem,

(24) $$\lim_{n\to\infty} \int_0^\infty \frac{dx}{\varphi_n(x, -s)^2} = \int_0^\infty \frac{dx}{\varphi_0(x, -s)^2}, \qquad s > 0.$$

This proves $h_n(s) \to h_0(s)$.

(iii)$\Rightarrow$(i). The family of left-continuous, non-decreasing functions $\{\operatorname{arctg} m_n(x)\}_{n=1}^\infty$ is uniformly bounded and therefore we can choose an inextensible measure $dm^*(x)$ and a subsequence $\{m_{n_j}\}$ of an arbitrary subsequence $\{m_{n_j}\} \subset \{m_n\}$ such that

(25) $$m_{n_j}(x) \to m^*(x) \quad \textit{at each continuity point } x \in [0, \infty) \textit{ of } m^*(x).$$

Since we have already proved that (i) implies (iii), it holds that $h_{n_j}(s)$ converges to $h^*(s)$, $h^*(s)$ being the characteristic function of $m^*(x)$. Of course $h^*(s) = h_0(s)$ by the assumption (iii). Hence $m^*(x) = m_0(x)$ by the uniqueness part of the above Krein's theorem. This completes our theorem.

3. Asymptotic theorems of Krein's correspondence.

A function $R(x)$ defined on $[0, \infty)$ is called *regular varying at* $\infty[0]$ *with exponent* ρ $(-\infty < \rho < \infty)$ if

(26) $$\lim_{\substack{x\to\infty \\ [0]}} \frac{R(\lambda x)}{R(x)} = \lambda^\rho, \qquad \lambda > 0.$$

Especially when $L(x)$ varies regularly with exponent zero, $L(x)$ is said to *vary slowly* (cf. [4] p. 275). In case ρ is finite, $R(x)$ varies regularly with exponent ρ if and only if there exists a slowly varying function $L(x)$ such that

(27) $$R(x) = x^\rho L(x).$$

⇨ *Recognition Result*

- Sample: *A sample of Math Journal digitized using InftyReader*

Section 2

Toward Rich DML

Digitization of Math Journals

Different levels:

- Level 1: Scanned images of papers
  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

Digitization of Math Journals

Different levels:

- Level 1: Bitmap images of printed materials.
  e.g. GIF, TIFF
- Level 2:
  e.g. PDF w

Infty : Level 1 → Level 3

- 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

Process Flow of Digitization

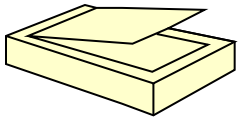


Image File (TIF)

PDF

Texts

Layout Analysis : Segmentation of Areas (Text, Table, Figure)

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

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

Outputs

XML

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

Layout Analysis

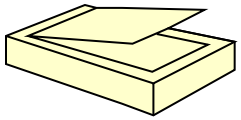
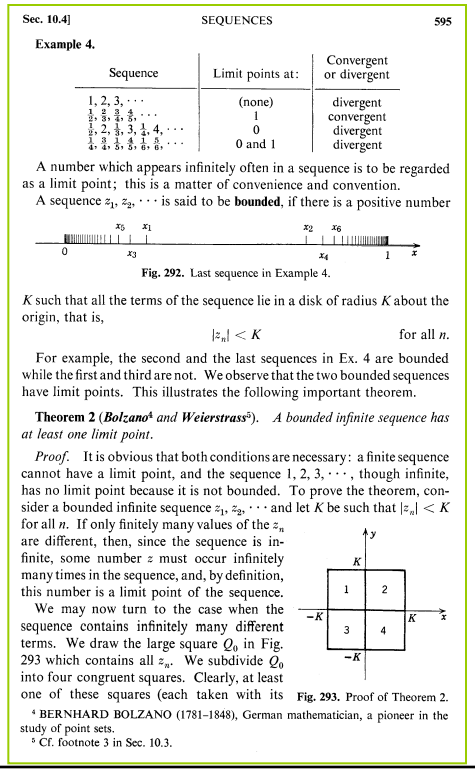


Image File (TIF)

PDF

(Pre processing)

Segmentation of Areas (Text, Table, Figure)



Sec. 10.4] SEQUENCES 595

Example 4.

Sequence	Limit points at:	Convergent or divergent
$1, 2, 3, \cdots$	(none)	divergent
$\frac{1}{2}, \frac{2}{3}, \frac{3}{4}, \frac{4}{5}, \cdots$	1	convergent
$\frac{1}{2}, 2, \frac{1}{3}, 3, \frac{1}{4}, 4, \cdots$	0	divergent
$\frac{1}{4}, \frac{3}{4}, \frac{1}{5}, \frac{3}{5}, \frac{1}{6}, \frac{5}{6}, \cdots$	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 $z_1, z_2, \cdots$ is said to be **bounded**, if there is a positive number

Fig. 292. Last sequence in Example 4.

K such that all the terms of the sequence lie in a disk of radius K about the origin, that is,

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

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 (***Bolzano***[4] *and* ***Weierstrass***[5]). *A bounded infinite sequence has at least one limit point.*

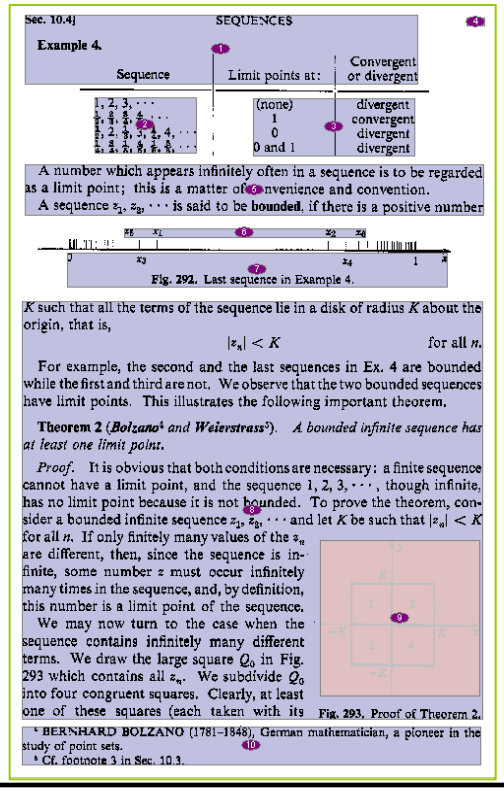
Proof. It is obvious that both conditions are necessary: a finite sequence cannot have a limit point, and the sequence $1, 2, 3, \cdots$, though infinite, has no limit point because it is not bounded. To prove the theorem, consider a bounded infinite sequence $z_1, z_2, \cdots$ and let K be such that $|z_n| < K$ 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. 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 Analysis

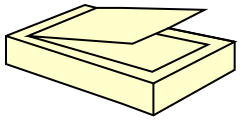


Image File (TIF) ⇦ PDF

Segmentation of Areas ⇨ Table Analysis

Example 4.

Sequence	Limit points at:	Convergent or divergent
$1, 2, 3, \cdots$	(none)	divergent
$\frac{1}{2}, \frac{2}{3}, \frac{3}{4}, \frac{4}{5}, \cdots$	1	convergent
$\frac{1}{2}, 2, \frac{1}{3}, 3, \frac{1}{4}, 4, \cdots$	0	divergent
$\frac{1}{4}, \frac{3}{4}, \frac{1}{5}, \frac{4}{5}, \frac{1}{6}, \frac{5}{6}, \cdots$	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 $z_1, z_2, \cdots$ is said to be **bounded**, if there is a positive number

Process Flow of Digitization

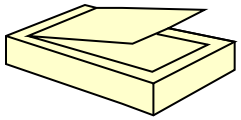


Image File (TIF)

PDF

Texts

Layout Analysis : Segmentation of Areas (Text, Table, Figure)

Line Segmentation

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

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

Outputs

XML

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

Line Segmentation (Sample)

so for r large enough, $J(z) \leqq C_2 r^{\varrho}$ for
$\overline{\lim}_{r \to \infty} J(rz)/r^{\varrho}$ and $j^*(z) = \overline{\lim}_{z' \to z} j(z')$, whi
and positively homogeneous of order

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

Line Segmentation (Sample)

so for r large enough, $J(z) \leqq C_2 r^{\varrho}$ for

$\overline{\lim}_{r \to \infty} J(rz)/r^{\varrho}$ and $j^*(z) = \overline{\lim}_{z' \to z} j(z')$, whi

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-

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

A Method of Line Segmentation

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

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

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 $\lim_{z \to 1} f(z) = 6$. (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

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

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 $\lim_{z \to 1} f(z) = 6$. (Note that the limit is established when some formula is found for δ as a function of ϵ.)

Process Flow of Digitization

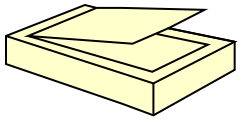


Image File (TIF)

PDF

Texts

Layout Analysis : Segmentation of Areas (Text, Table, Figure)

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

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

Outputs

XML

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

Math/Text Segmentation

Number of characters in Math area is about 20% of all the characters in pure math journals.

No math. structure

We assert that X is torsion-free. Indeed, if X is not torsion-free then it has a direct summand $C(p^k)$, $1 \leqq k < \infty$ ([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

Math/Text Segmentation

- **Recognition of Ordinary Texts & Math/Text Area Segmentation**
  = *Simultaneous Process using DP*

1. Combination of different OCR engines
2. Score using relative position check

Current version:

Infty + Two commercial OCRs (Toshiba + Media Drive)

New version: FineReader engine will be added.

Methods

<u>Method 1</u> (Recognition of words)

Niihama-gun,

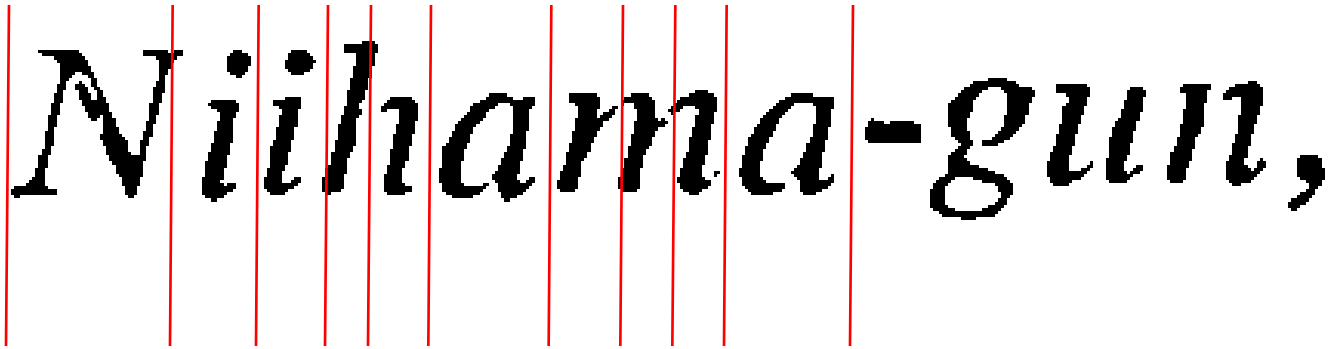
F : *N ü h a m a - g u n ,*

E : *N i i h a r n a - g u n ,*

I : *N i i / ı a m a - g u n ,*

Methods

Method 1.



F : *N ü h a m a - g u n ,*

E : *N i i h a r n a - g u n ,*

I : *N i i / ı a m a - g u n ,*

Methods

Method 1.

Result *N i i h a m a g u n ,*

F : *N ü h a m a - g u n ,*

E : *N i i h a r n a - g u n ,*

I : *N i i / ı a m a - g u n ,*

Methods

<u>Method 2</u> (Use character sizes and positions)

Consider the metric ϱ defined in Ω by

$$\varrho(z)\,|dz| = \frac{|du + i * du|}{\Theta(c)} \quad \text{if} \quad z \in l(c),\ -\infty < c < \infty.$$

It is not difficult to show (cf., e.g., Ohtsuka [7]) that, if $\Gamma(a, b) \neq \emptyset$, th

```
Consider the metric Q defined in O by
¥du + i*du¥ .. _. .
= J Q^T L if Z&l(c), -oo.<c<oo.
It is not difficult to show (cf., e.g., Ohtsuka [7]) that,|
```

Methods

<u>Method 2</u> (Use character sizes and positions)

Consider the metric ϱ defined in Ω by

$$\varrho(z)\,|dz| = \frac{|du + i * du|}{\Theta(c)} \quad \text{if} \quad z \in l(c),\ -\infty < c < \infty,$$

It is not difficult to show (cf., e.g., Ohtsuka [7]) that, if $\Gamma(a, b) \neq \varnothing$, th

```
Consider the metric Q defined in O by
¥du + i*du¥ .. _. .
= J Q^T L if Z&l(c), -oo.<c<oo.
It is not difficult to show (cf., e.g., Ohtsuka [7]) that,|
```

Process Flow of Digitization

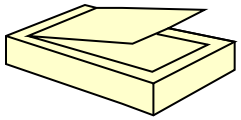


Image File (TIF)

PDF

Texts

Layout Analysis : Segmentation of Areas (Text, Table, Figure)

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

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

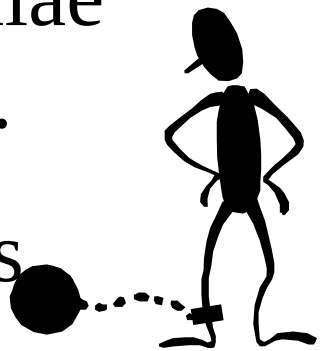
Outputs

XML

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

Formulae Recognition

- Recognition of *Variety of Rare Symbols*
- Distinction of Fonts.
  (Italic, Bold, Bbb, Caligraphic,etc.)
- Segmentation of *Touched/Broken* characters in Math Area is still a difficult problem.
- *Stable Structure Analysis* of math formulae against the miss-recognition of characters.
- Distinction of *Noises* and Small symbols



Process Flow of Digitization

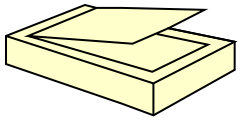


Image File (TIF)

PDF

Texts

Layout Analysis : Segmentation of Areas (Text, Table, Figure)

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

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

Outputs

XML

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

Document Structure Analysis

- Detection of :

  Title, Autor, Section, Subsection, Itemization, BibItem, Theorem, Lemma, etc.

  - A Naïve method:

    Line classification using the combination features such as:
    Character size, Font Information (Bold, Italic, Small Capital), Keywords, Indentation, Starting with Numbers or Special pattern (e.g. "[Num]"), etc.

  - Stronger method is required in actual digitization.

- Hyperlink inside document.

Section 3

Formulae Recognition by Infty

Formulae recognition

• R. Anderson, Syntax directed recognition of hand-printed two dimensional mathematcs.
Interactive System for Experimental Applied Mathematics, M.Klerer and J. Reinfelds, Eds, Academic Press, 1968, pp. 436-459

• M. Okamoto and H. Twaakyondo, Structure analysis and recognition of mathematical expressions, 3rd ICDAR, 1995, Montreal, (1995), 430--437.

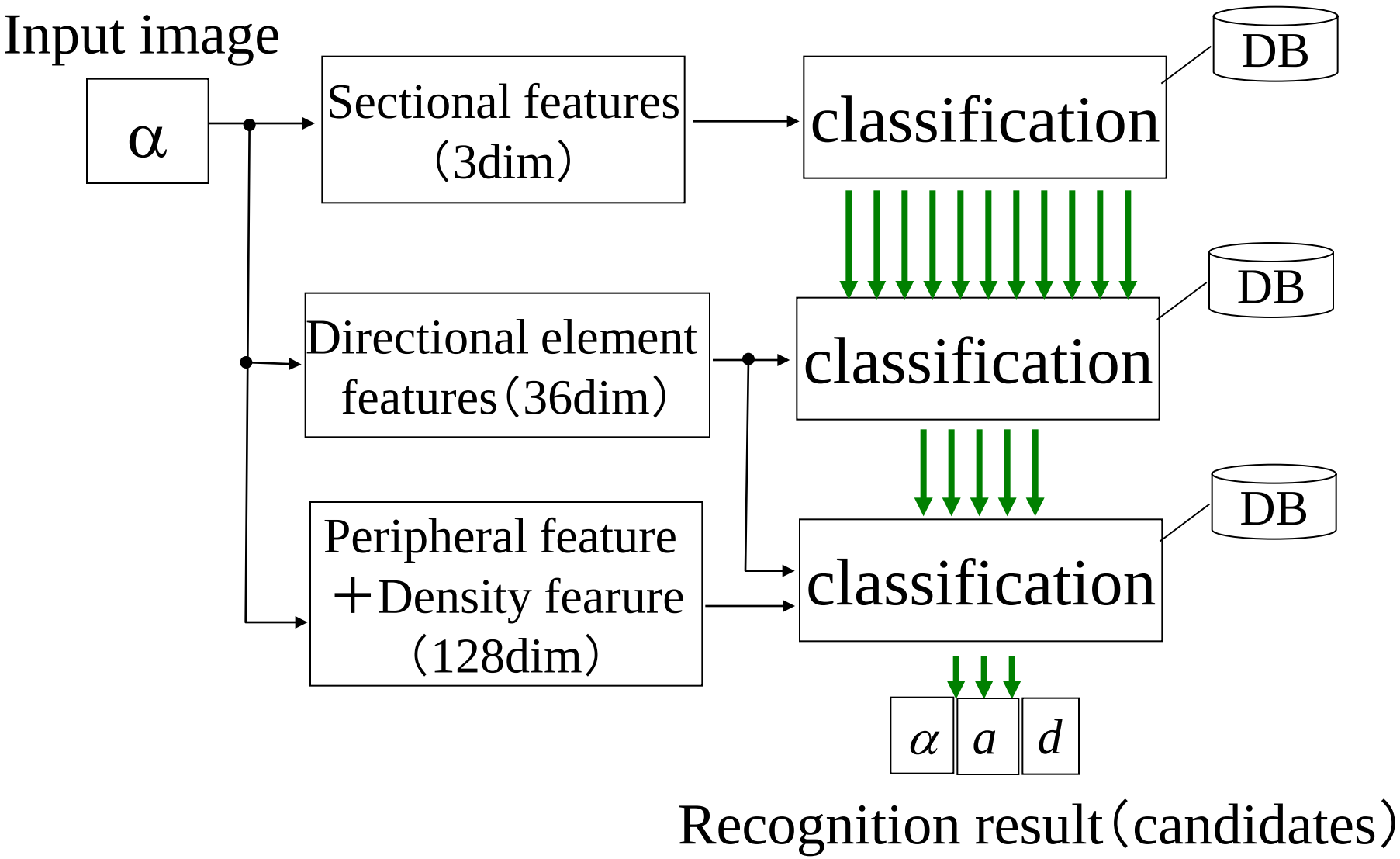
• R. J. Fateman, T. Tokuyasu, B. P. Berman and N.Mitchell, Optical Character Recognition and Parsing of Typeset Mathematics, Journal of Visual Communication and Image Representation vol.7, no.1, (1996), 2--15.

• Y. Eto and M.Suzuki, Mathematical formula recognition using virtual link network, 6th ICDAR, 2001, Seattle, IEEE Computer Society Press , 430--437

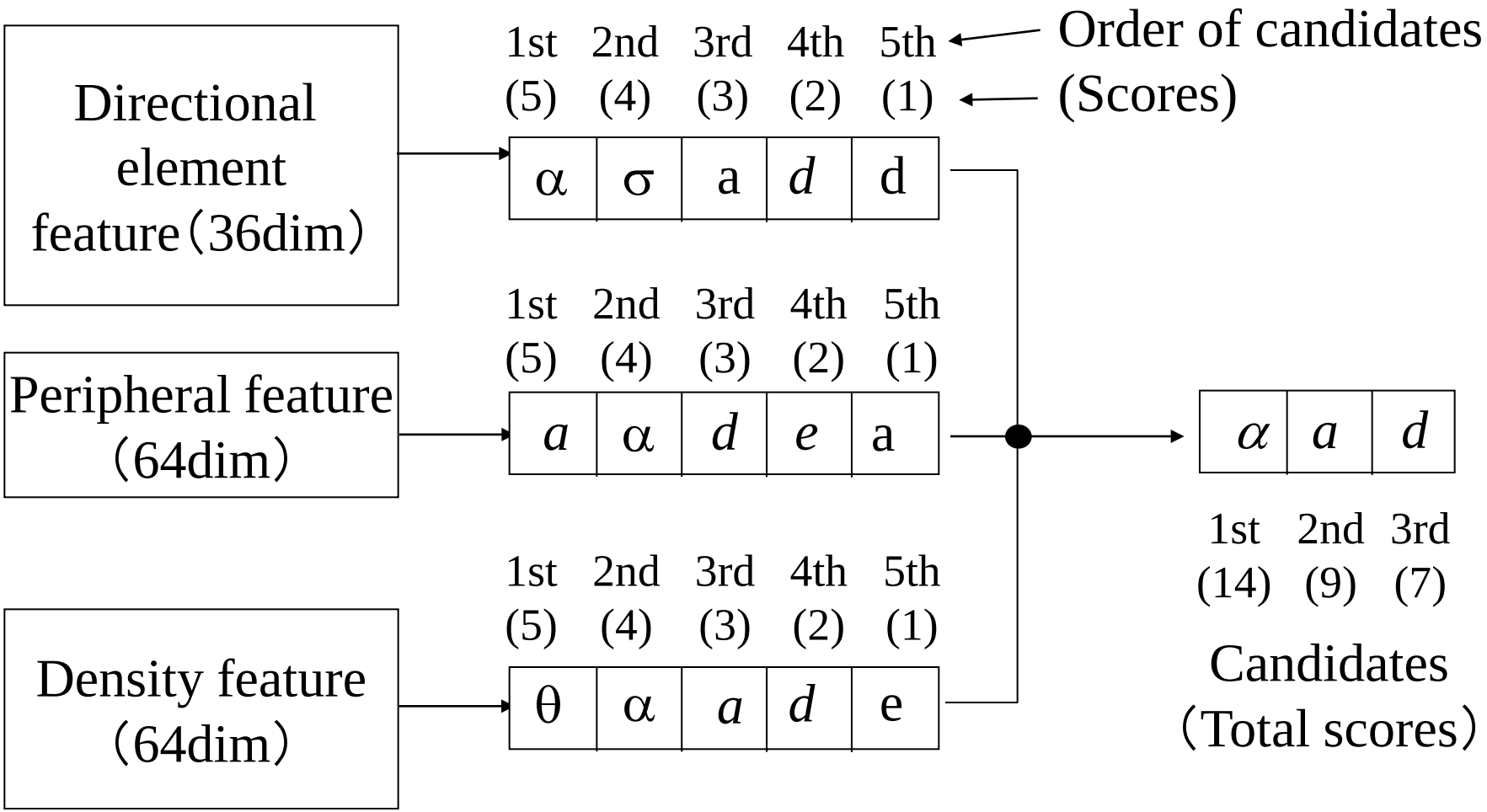
Infty OCR engine

- 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

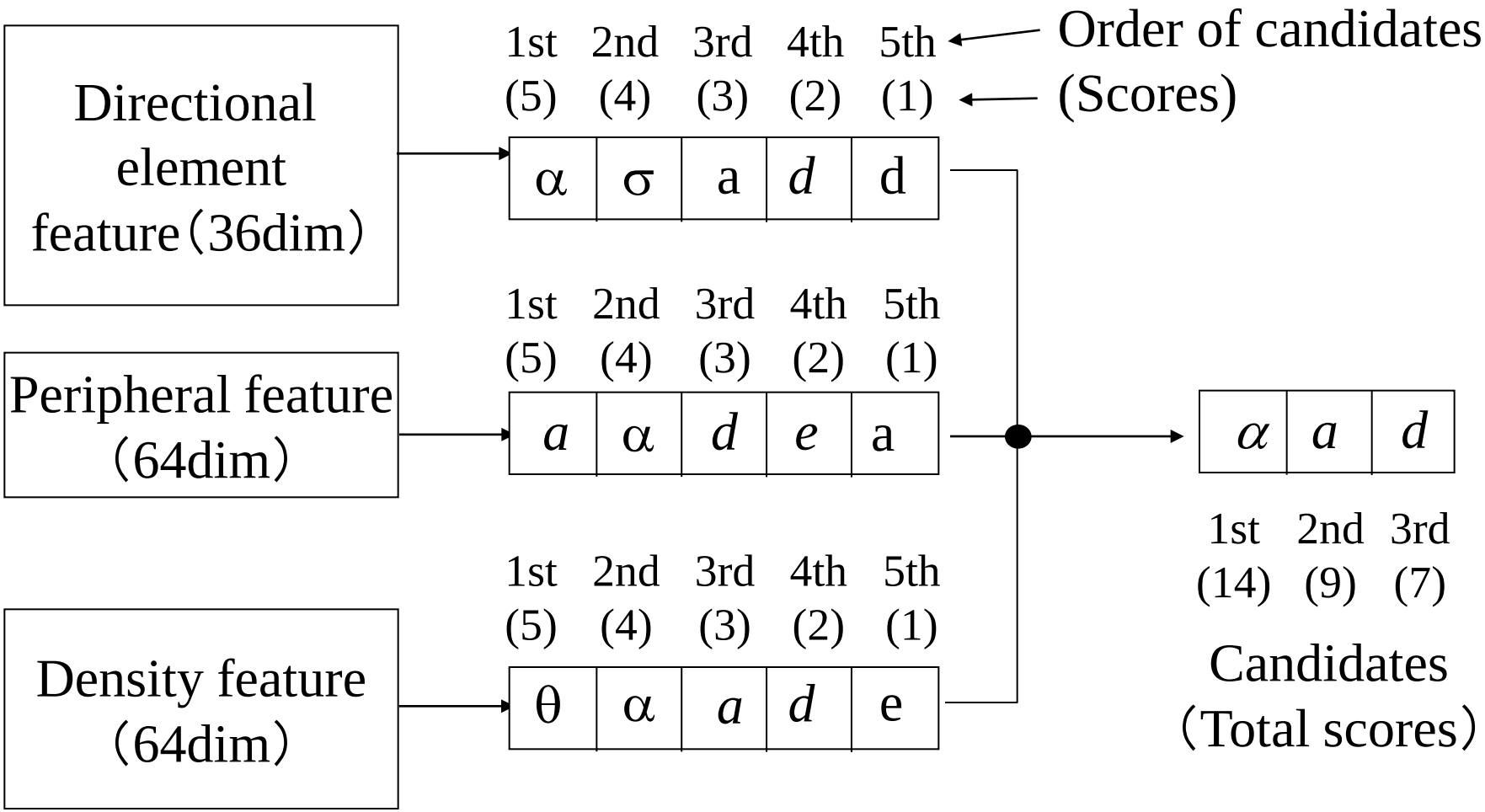
3 step classifications



Voting method



Voting method

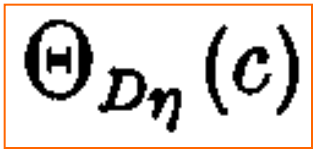


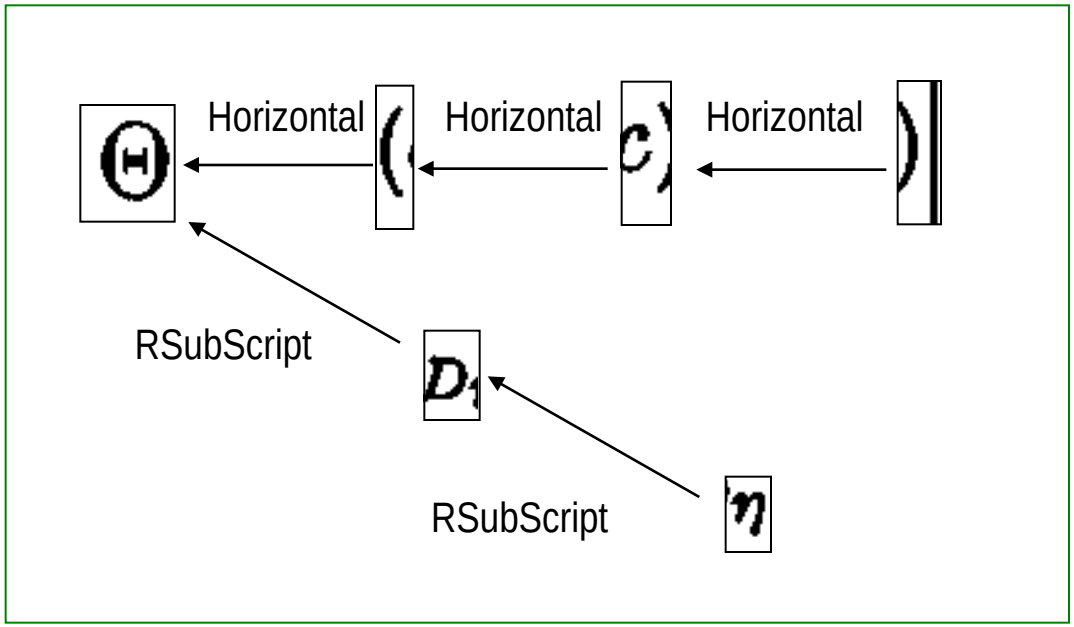
Voting → *Normalization of the score of symbol recognition*

Structure Analysis of Formulae

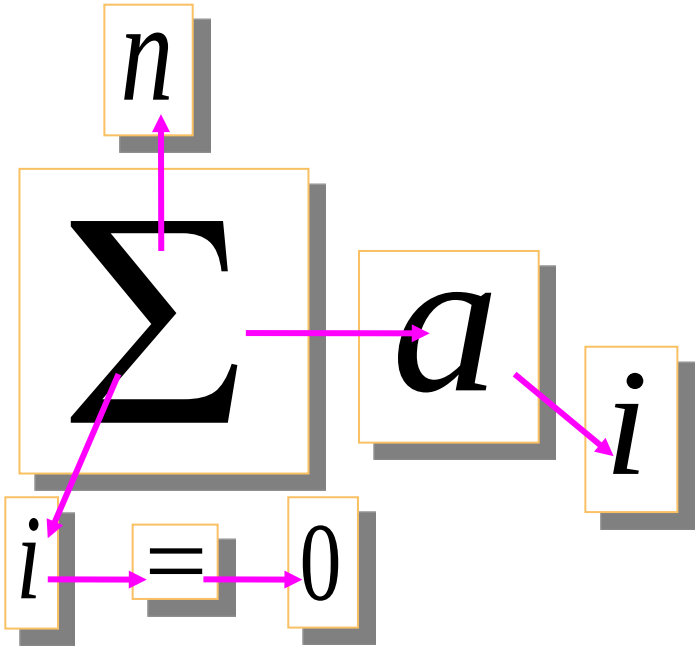
Output（Tree Structure）

Input (image)





Structure Analysis of Formulae



Structure Analysis of Formulae

- Some difficult cases :

Collapsing of quotient spaces of $SO(n)\backslash SL(n,\boldsymbol{R})$ at infinity 203

$$= -C \log \left\{ \Delta_1(x)^{\alpha_1 - \alpha_n} \times \prod_{k=2}^{n-1} \left(\frac{\Delta_k(x)}{\Delta_{k-1}(x)} \right)^{\alpha_k - \alpha_n} \right\}$$

$$= -C \log \qquad \Delta_k(x)^{\alpha_k - \alpha_{k+1}}$$

$$= C \log \left(\prod_{k=1}^{n-1} \Delta_k(x)^{\alpha_{k+1} - \alpha_k} \right) \qquad \square$$

Structure Analysis of Formulae

- Some difficult ca

Collapsing of quotient space

$$= -C \log \left\{ \Delta_1(x) \right.$$

$$= -C \log \quad \Delta_k(x)^{\alpha_k - \alpha_{k+1}}$$

$$= C \log \left(\prod_{k=1}^{n-1} \Delta_k(x)^{\alpha_{k+1} - \alpha_k} \right) \qquad \square$$

$$\Delta_k(x)^{\alpha_k - \alpha_{k+1}}$$

Structure Analysis of Formulae

- Some diffi

Collapsing of quotie

$$= C \log \left(\prod_{k=1}^{n-1} \Delta_k(x)^{\alpha_{k+1} - \alpha_k} \right)$$

$$= -C \log$$

$$= -C \log \quad \Delta_k(x)^{\alpha}$$

$$= C \log \left(\prod_{k=1}^{n-1} \Delta_k(x)^{\alpha_{k+1} - \alpha_k} \right) \qquad \square$$

Structure Analysis of Formulae

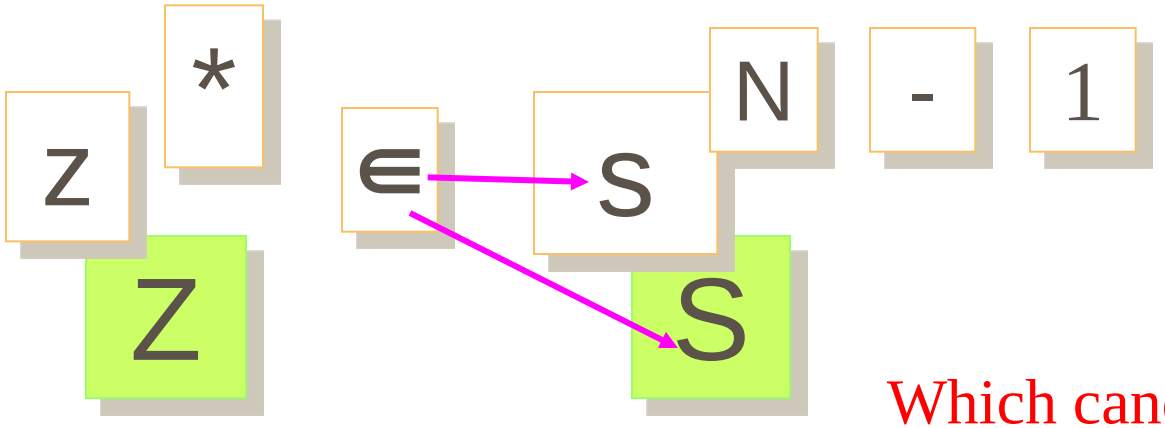
- Link possibilities :

$$= c_t^*(25)$$

Structure Analysis of Formulae

- Similar characters :

$$z^* \in S^{N-1}$$

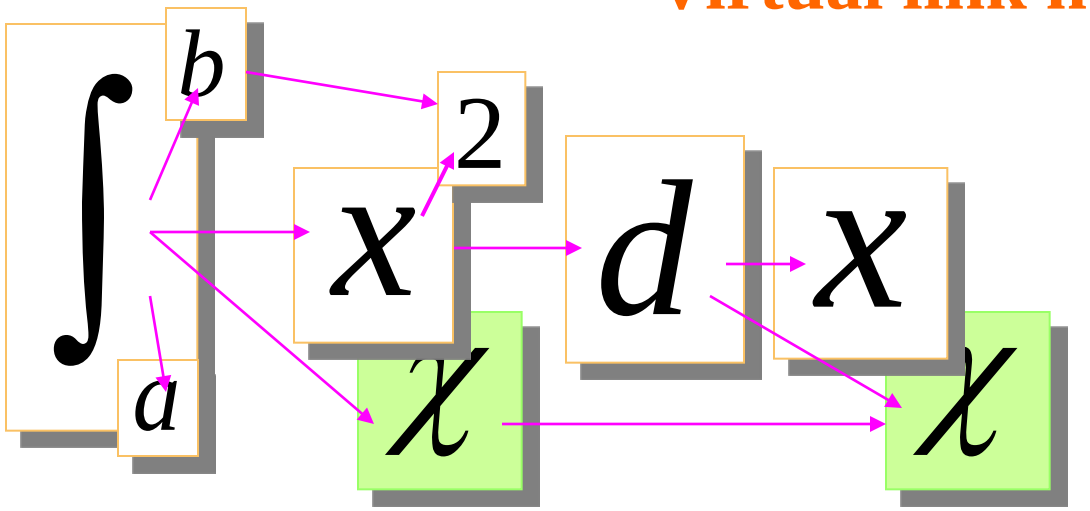


Which candidates are appropiriate?

Structure Analysis of Formulae

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

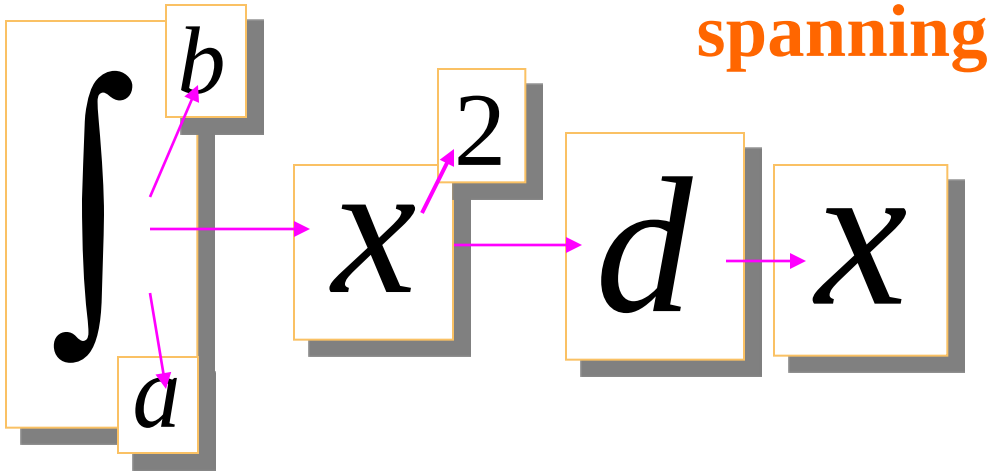
Virtual link network



Structure Analysis of Formulae

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

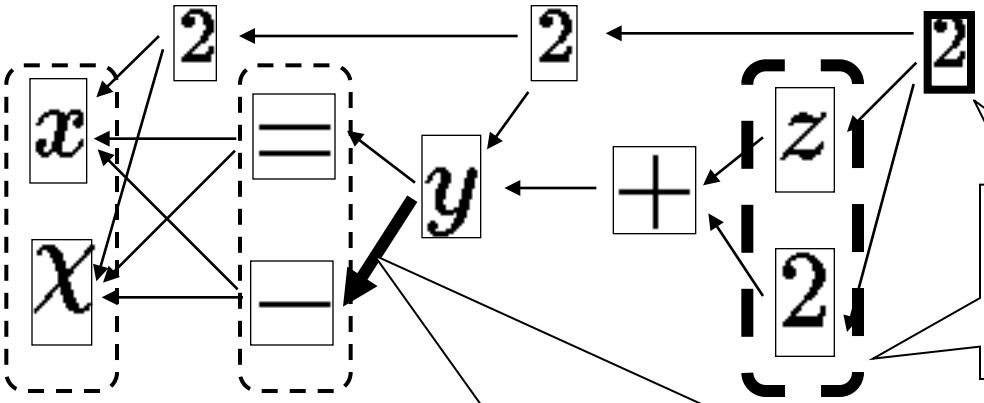
Search for correct spanning tree



Virtual link network

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

Virtual link network



Nodes =Candidates of character recognition

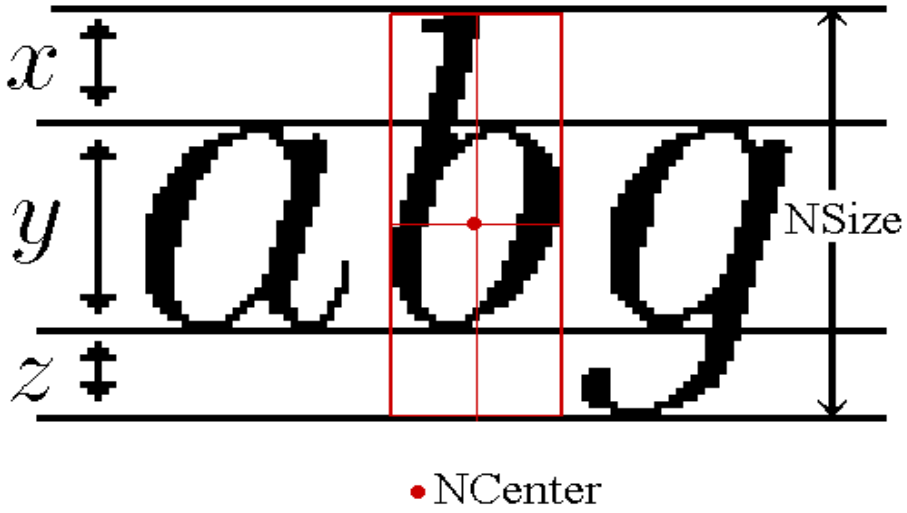
Each Link has a label and the link cost

Link: Horizontal, Upper, Under, Rsup, Rsub,Lsup, Lsub

Link Cost

Definitions of :

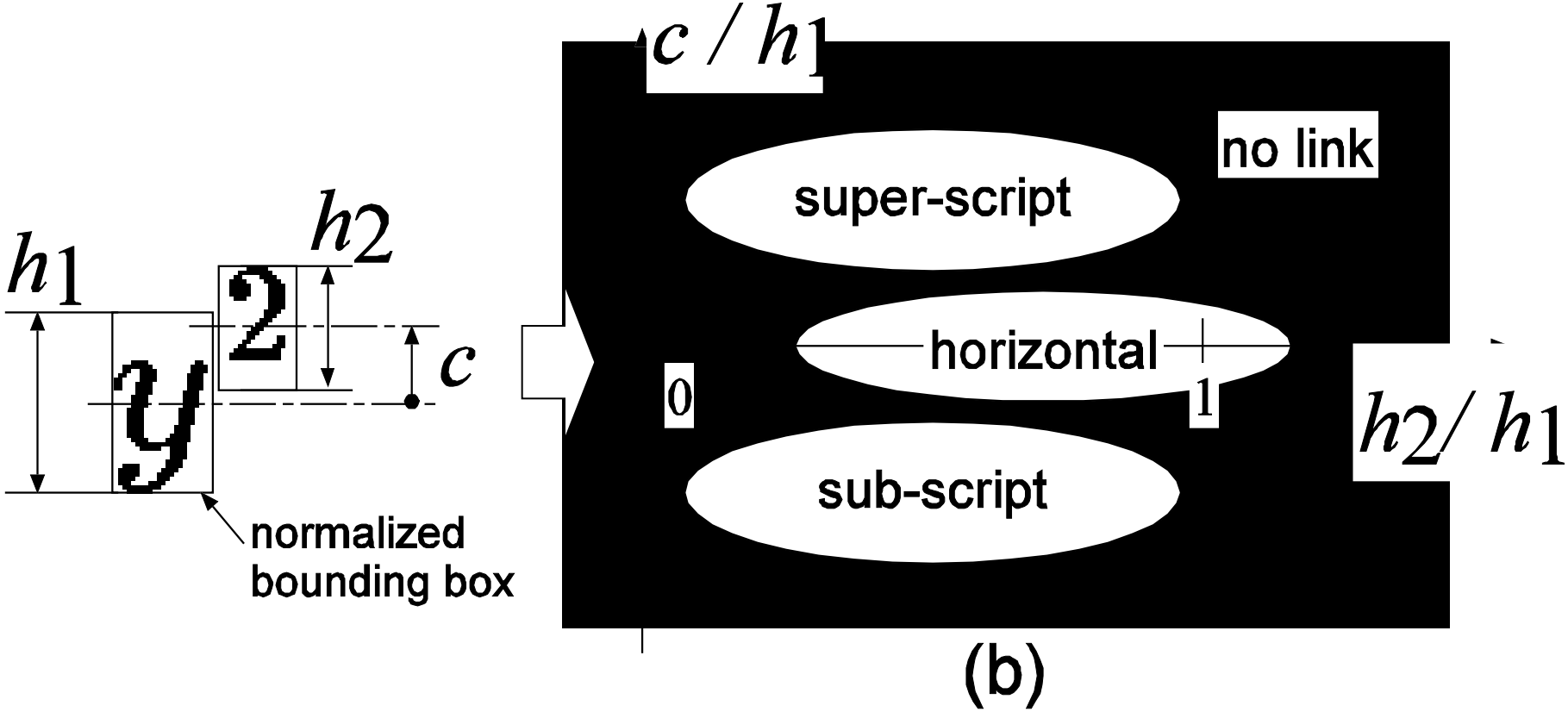
Normalized size (NSize) and
Normalized center (NCenter)



x:y:z = 28:51:21

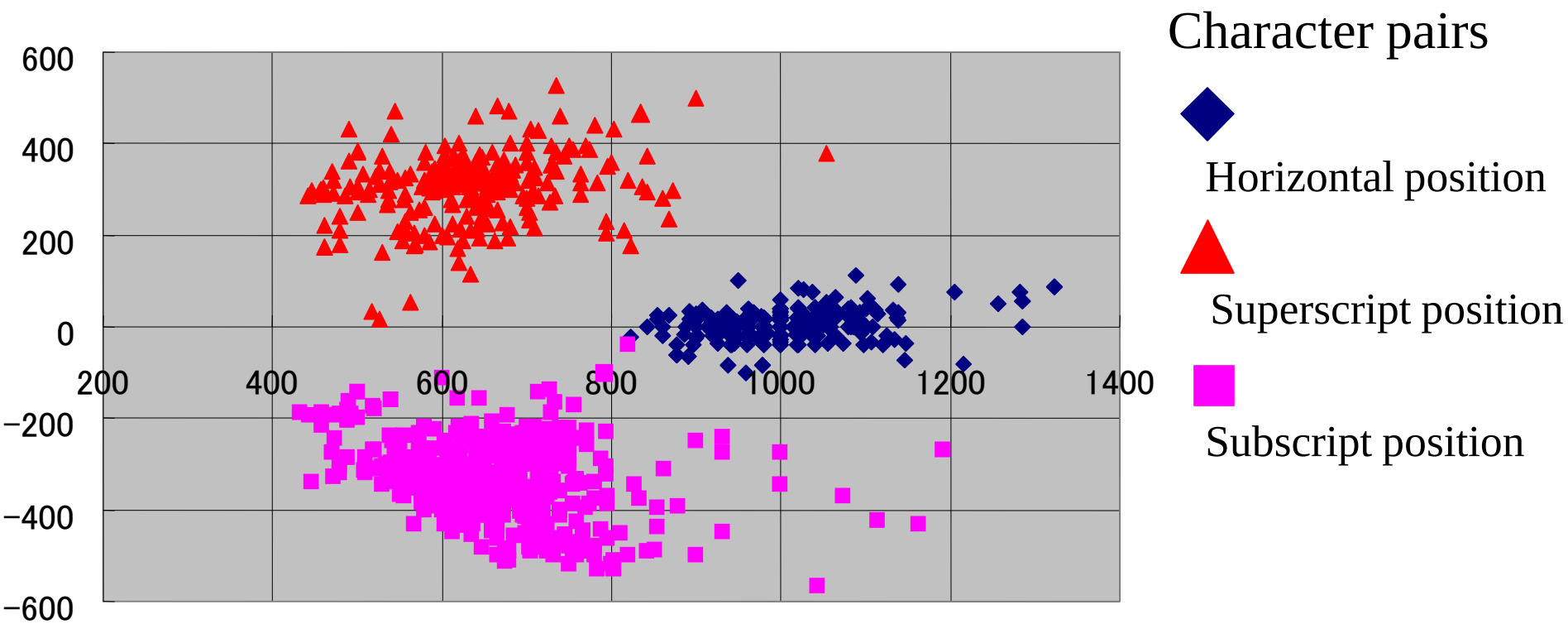
(dafault value)

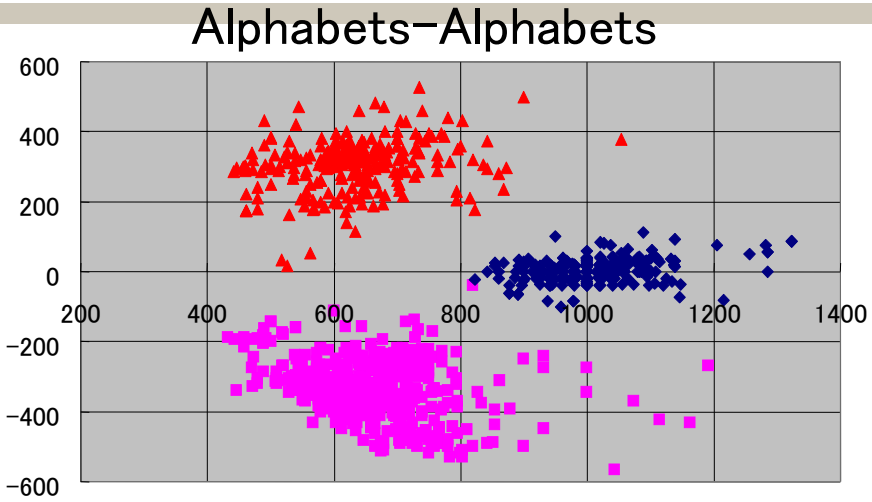
Link Cost

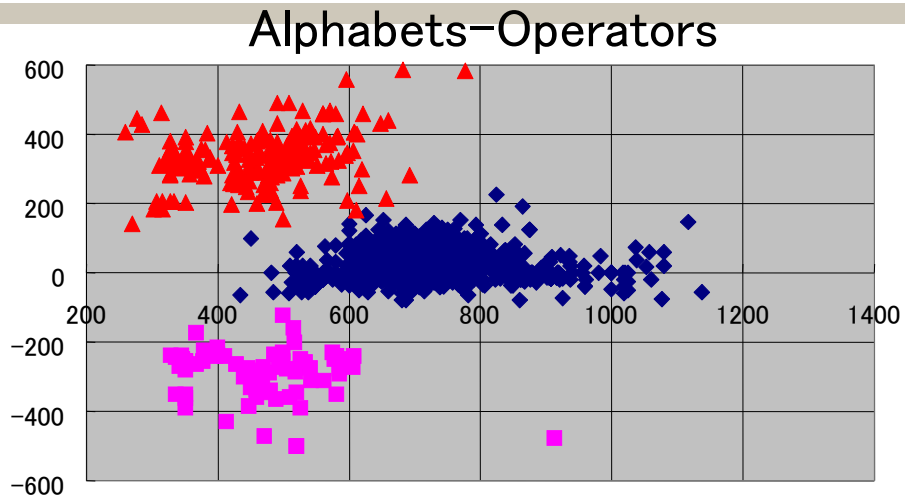

h1
h2
c
normalized
bounding box
c / h1
no link
super-script
horizontal
0
1
h2/ h1
sub-script
(b)

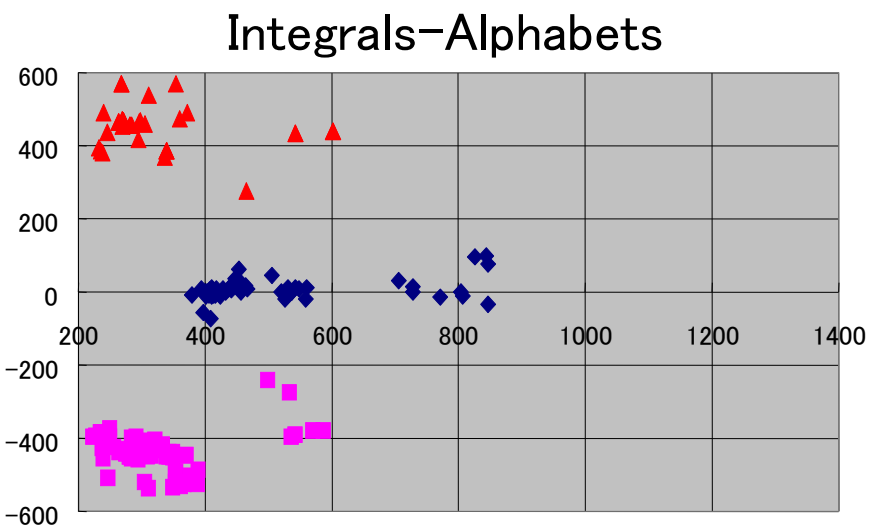
Link Cost

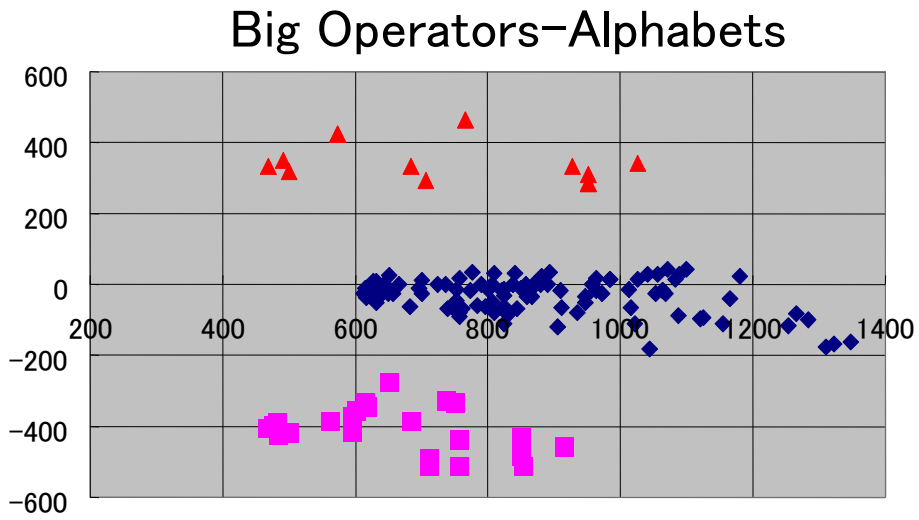
Distribution map in the (H,D)-plane




Alphabets-Alphabets
600
400
200
0
-200
-400
-600
200
400
600
800
1000
1200
1400

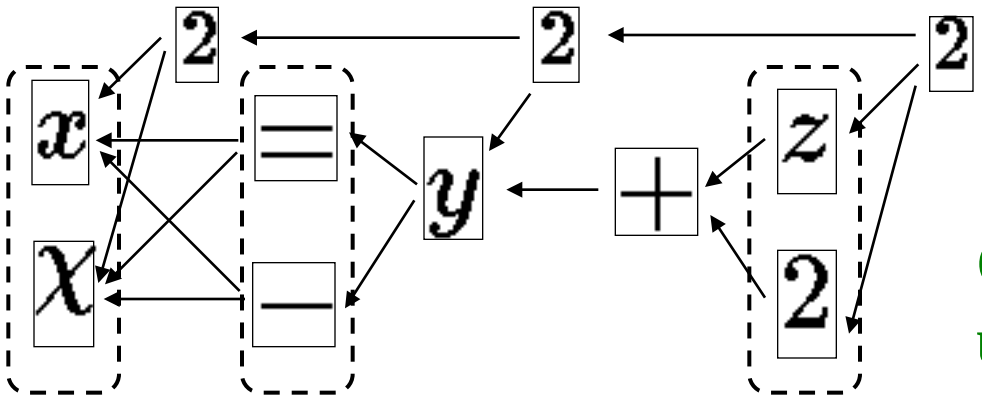

Alphabets-Operators
600
400
200
0
-200
-400
-600
200
400
600
800
1000
1200
1400


Integrals-Alphabets
600
400
200
0
-200
-400
-600
200
400
600
800
1000
1200
1400


Big Operators-Alphabets
600
400
200
0
-200
-400
-600
200
400
600
800
1000
1200
1400

Extraction of Structure Tree

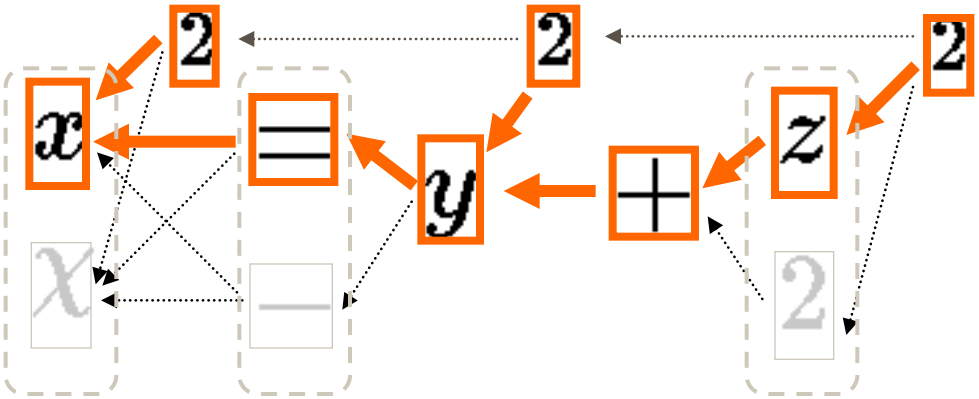
Network



Optimization under constraints

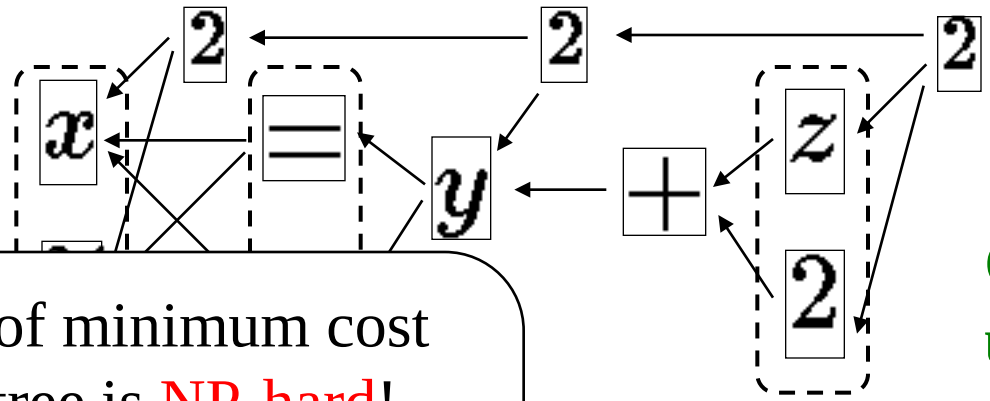
- Minimum total cost
- Link restrictions

Structure Tree



Extraction of Structure Tree

Network



Extraction of minimum cost spanning tree is NP-hard!

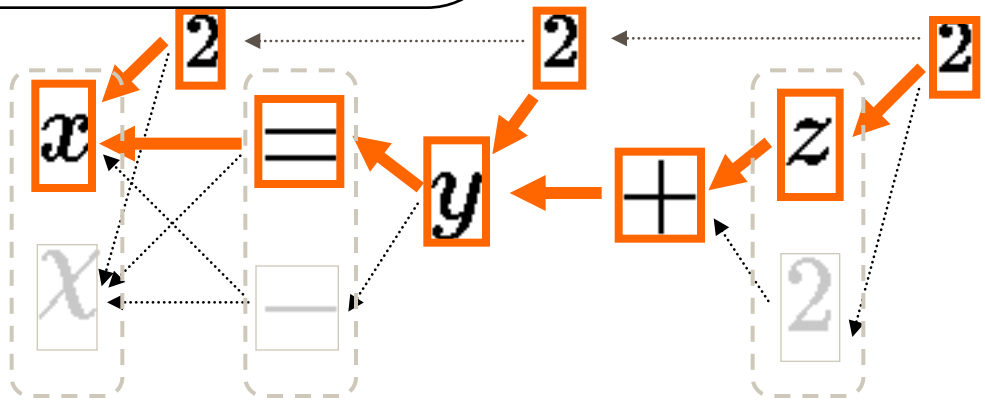
↓

Strategy of the current version: Beam search

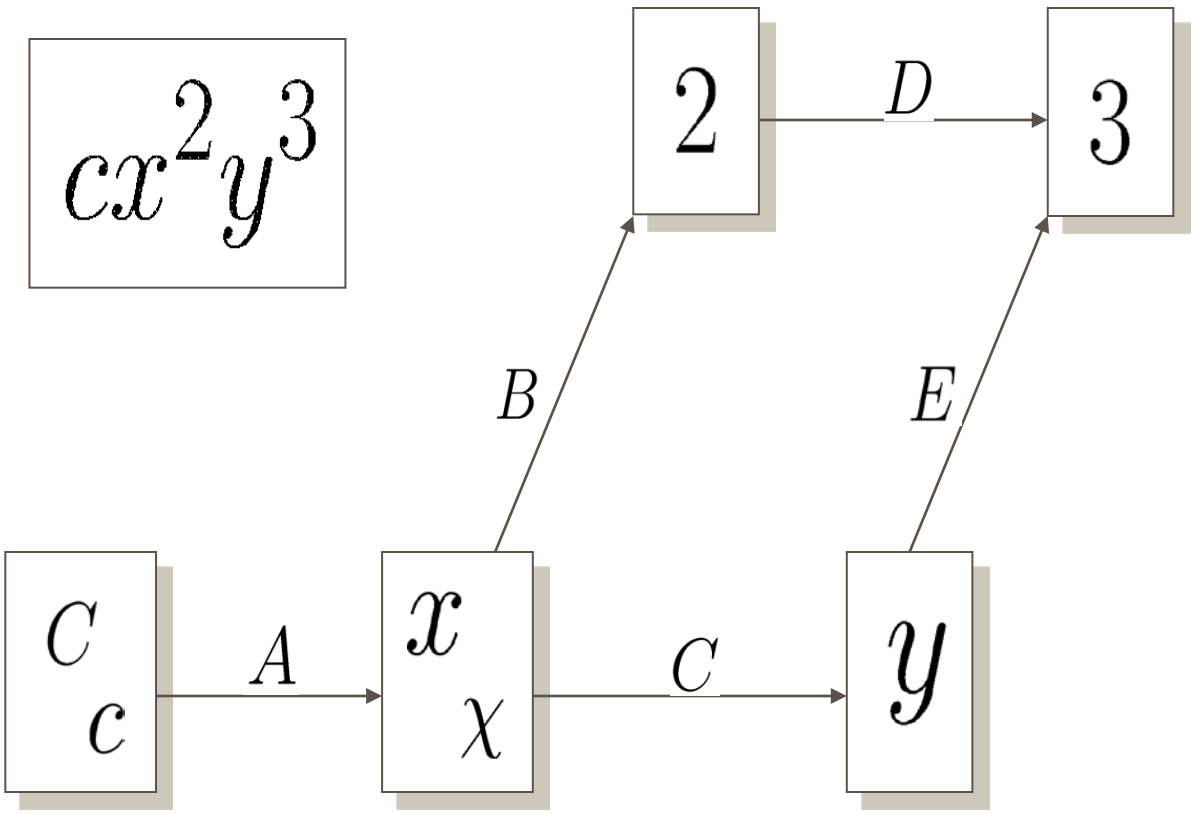
Optimization under constraints

- Minimum total cost
- Link restrictions

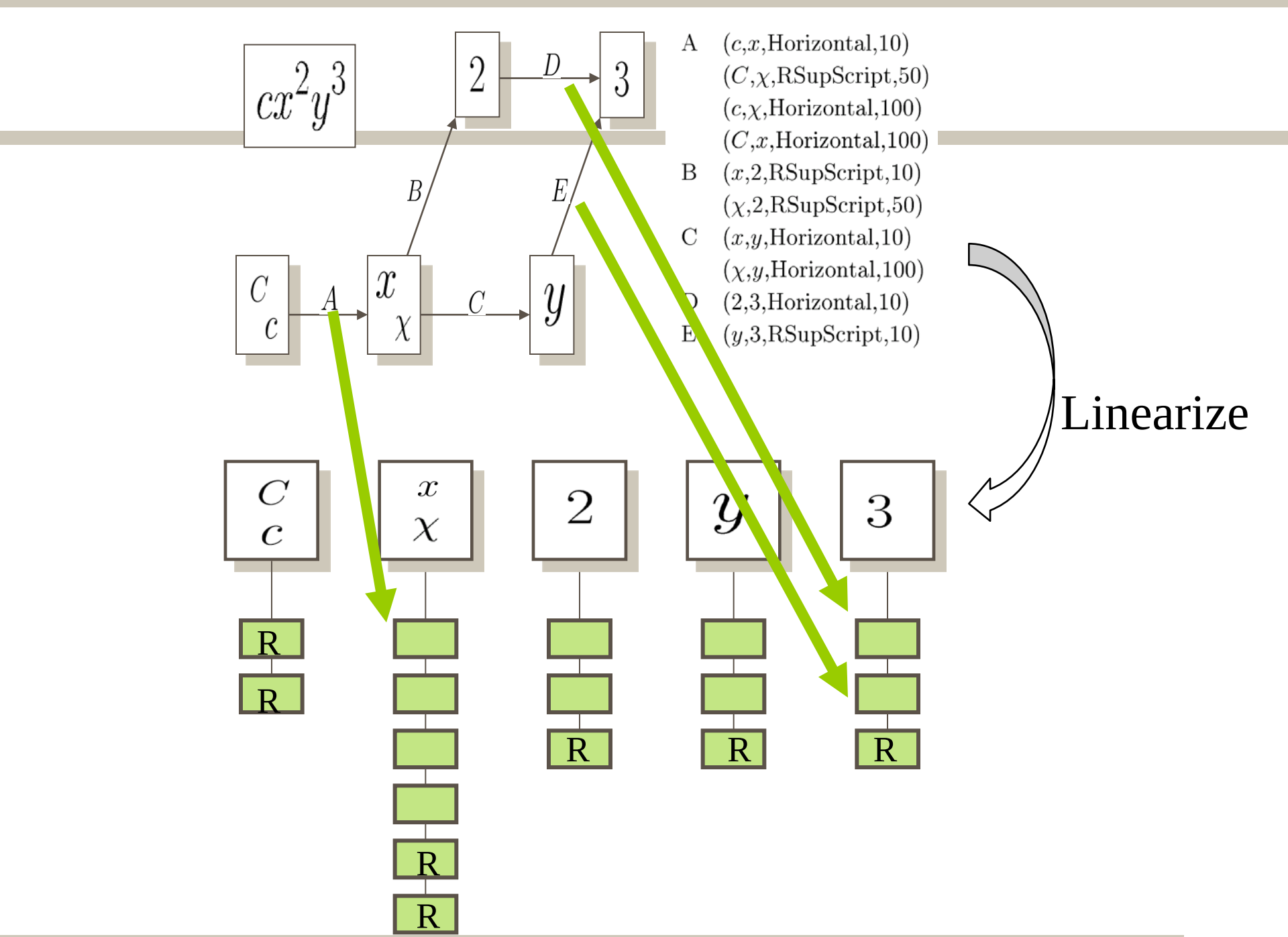
Tree

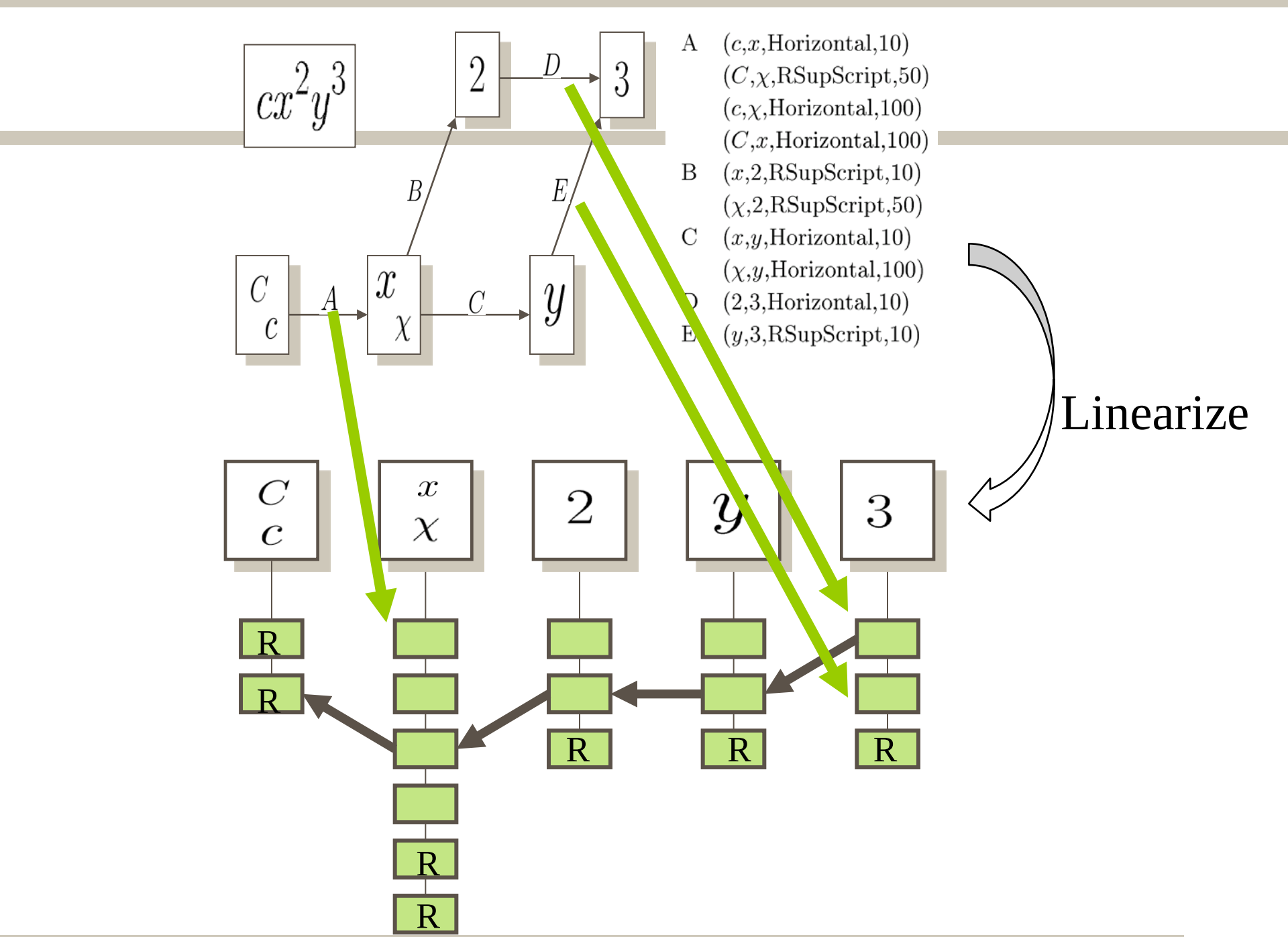


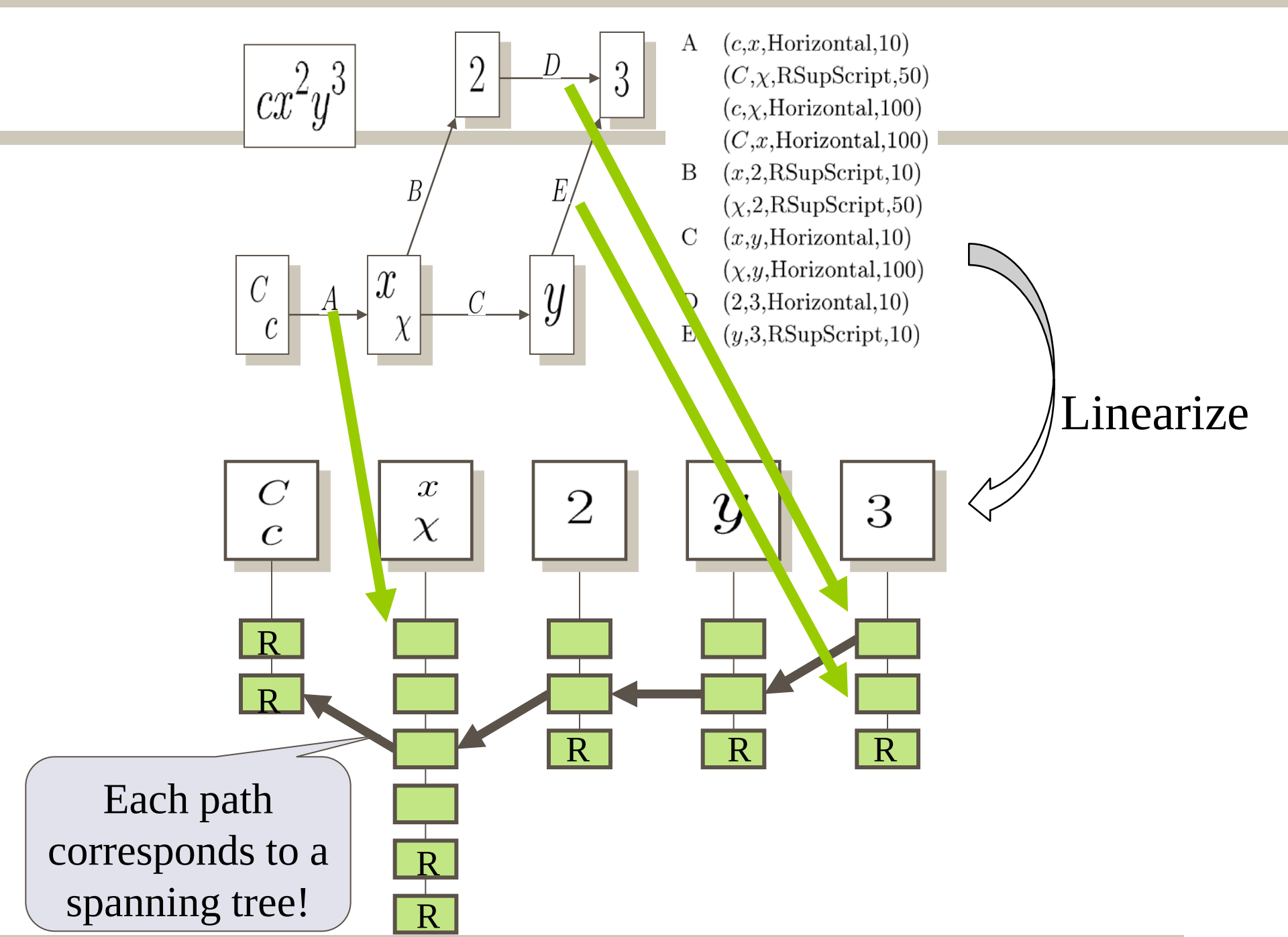
Extraction of Structure Tree



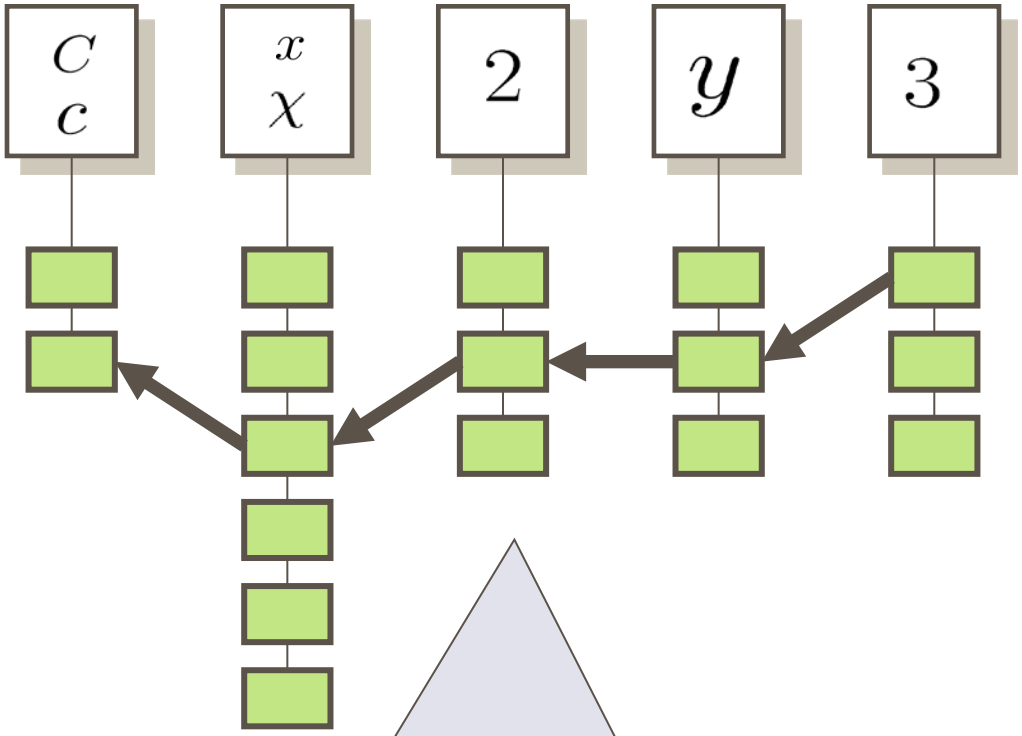
A	(c,x,Horizontal,10)
	(C,χ,RSupScript,50)
	(c,χ,Horizontal,100)
	(C,x,Horizontal,100)
B	(x,2,RSupScript,10)
	(χ,2,RSupScript,50)
C	(x,y,Horizontal,10)
	(χ,y,Horizontal,100)
D	(2,3,Horizontal,10)
E	(y,3,RSupScript,10)


cx^2y^3
2
D
3
B
E
C
c
A
x
χ
C
y
A (c,x,Horizontal,10)
(C,χ,RSupScript,50)
(c,χ,Horizontal,100)
(C,x,Horizontal,100)
B (x,2,RSupScript,10)
(χ,2,RSupScript,50)
C (x,y,Horizontal,10)
(χ,y,Horizontal,100)
D (2,3,Horizontal,10)
E (y,3,RSupScript,10)
Linearize
R


cx^2y^3
2
D
3
B
E
C
c
A
x
χ
C
y
A (c,x,Horizontal,10)
(C,χ,RSupScript,50)
(c,χ,Horizontal,100)
(C,x,Horizontal,100)
B (x,2,RSupScript,10)
(χ,2,RSupScript,50)
C (x,y,Horizontal,10)
(χ,y,Horizontal,100)
D (2,3,Horizontal,10)
E (y,3,RSupScript,10)
Linearize
C
c
x
χ
2
y
3
R
R
R
R
R
R
R


cx^2y^3
2
D
3
B
E
C
c
A
x
χ
C
y
A (c,x,Horizontal,10)
(C,χ,RSupScript,50)
(c,χ,Horizontal,100)
(C,x,Horizontal,100)
B (x,2,RSupScript,10)
(χ,2,RSupScript,50)
C (x,y,Horizontal,10)
(χ,y,Horizontal,100)
D (2,3,Horizontal,10)
E (y,3,RSupScript,10)
Linearize
C
c
x
χ
2
y
3
R
R
R
R
R
R
R
R
Each path corresponds to a spanning tree!

Search of spanning tree



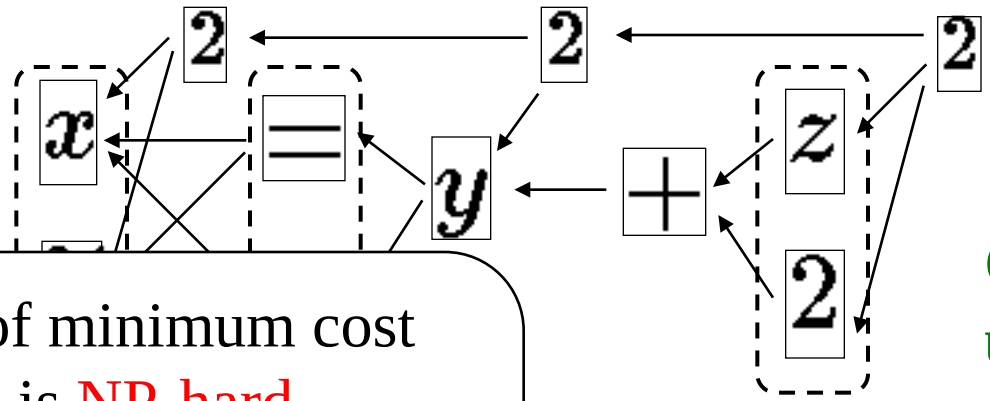
Each path corresponds to a spanning tree

Beam Search：

At each step, we hold a fixed number of paths (=Beam) with lowest costs, and use them at the next step.

Extraction of Structure Tree

Network

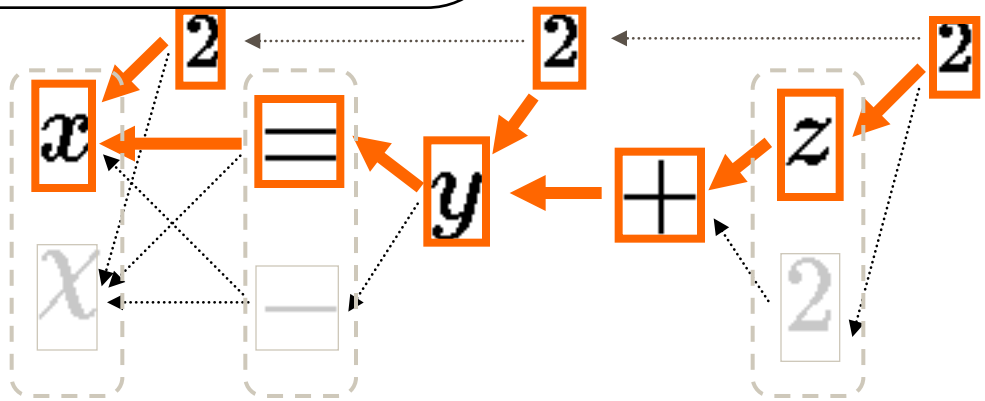


- Extraction of minimum cost spanning tree is NP-hard.
- Beam search fails sometimes to get optimal solution.

Optimization under constraints

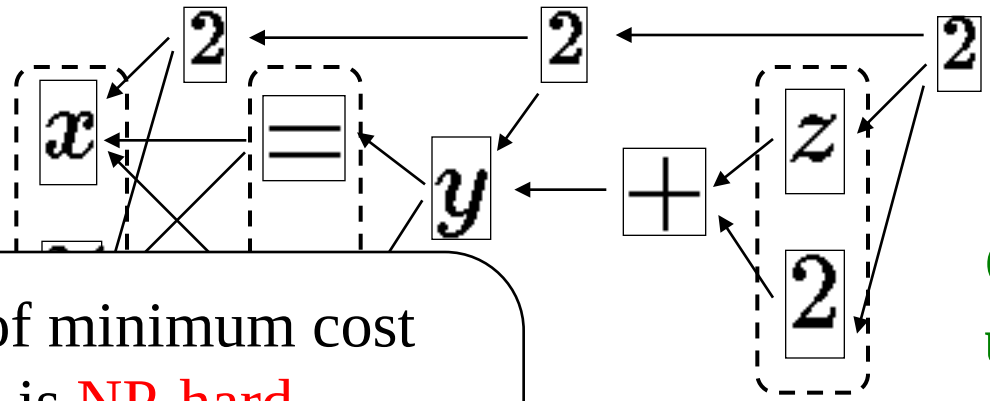
- Minimum total cost
- Link restrictions

Tree



Extraction of Structure Tree

Network

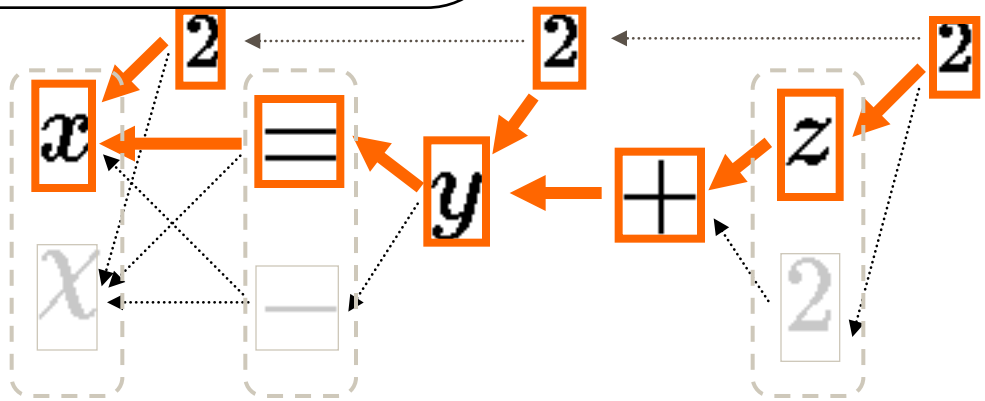


- Extraction of minimum cost spanning tree is NP-hard.
- Beam search fails sometimes to get optimal solution.
- Some other better strategy?

Optimization under constraints

- Minimum total cost
- Link restrictions

Tree



Section 4

Large Volume Recognition

Large Volume Digitization

- Retro-digitization of journals,
- Reproduction of old book/series of books,
- Translation to different languages,
- Braille transcription, DAISY talking book,
- etc.

Large Volume Digitization

- *Adaptive method* is efficient:

Get information from the target document:
- *Character features,*
- *Math formula parameters,*
- *Layout parameters, etc.*

(Directly) or After manual checking *(Semi-automatic)*

Recognition

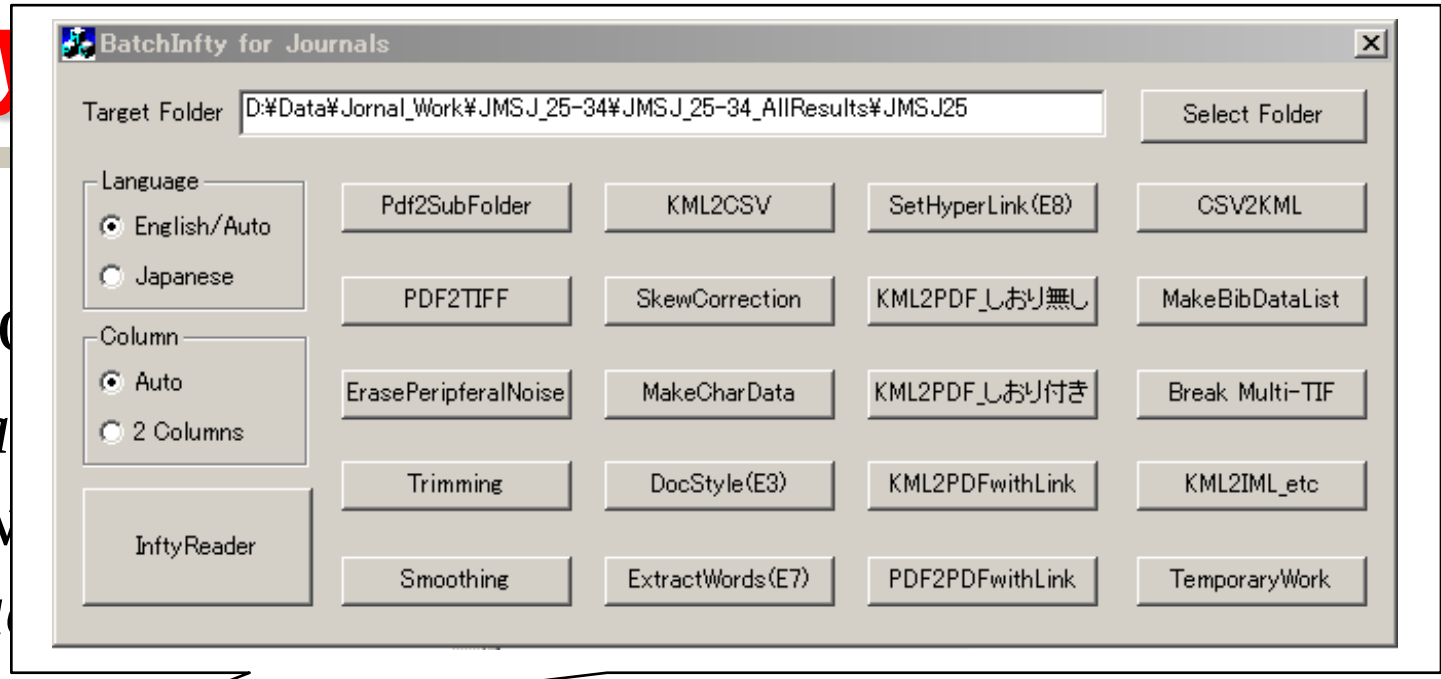
"InftySystem" for large scald digitization

- Applications:

1. *InftyReader* downloadable from our web site:
   http://www.sciaccess.net
2. *InftyReader Pro* (professional version)
3. *BatchInfty*
4. *CharImageManager*

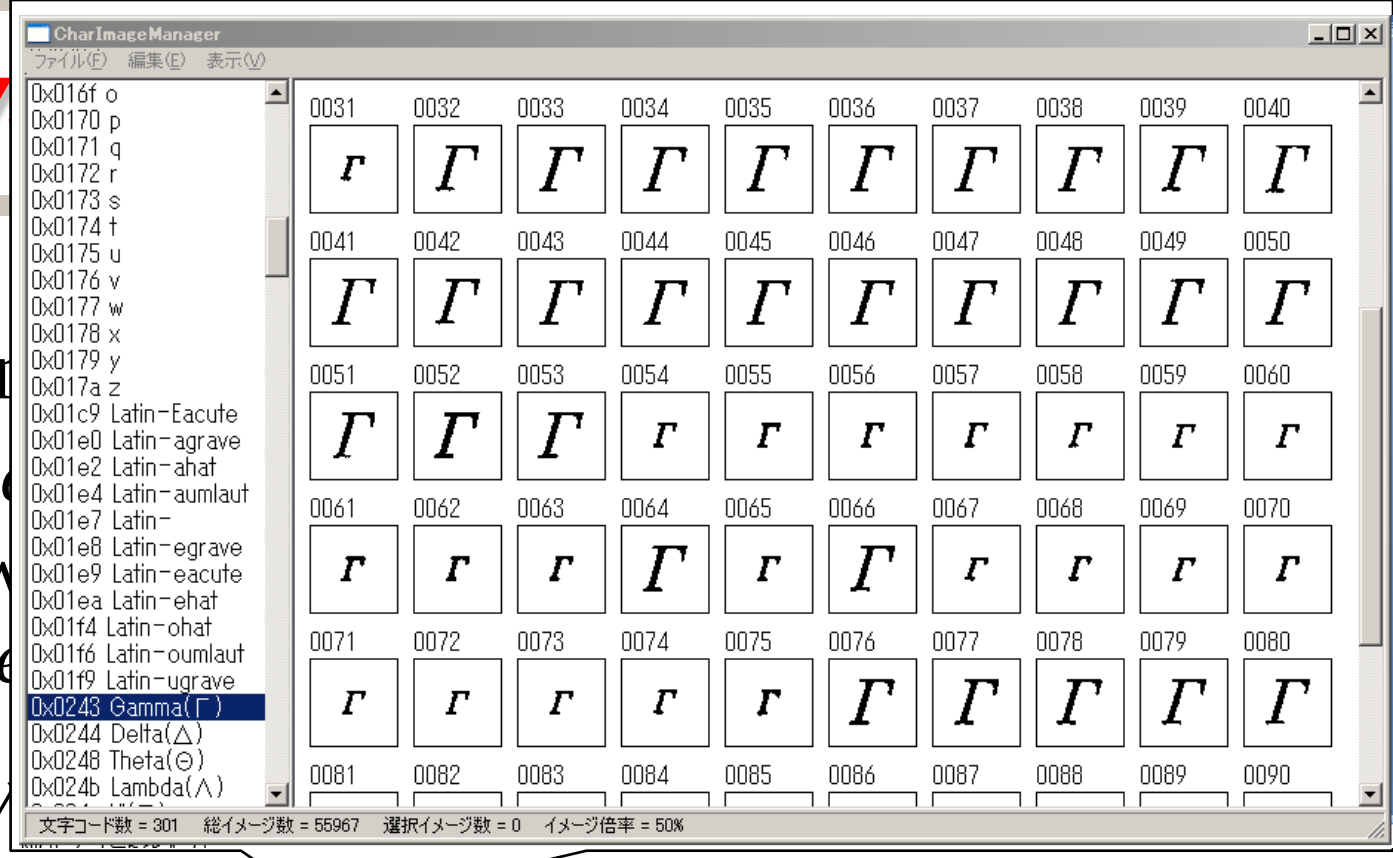
"Infty

- Applicati
  1. *InftyRea*
     http://w
  2. *InftyRea*
  3. *BatchInfty*
  4. *CharImageManager*



"Infty

- Application
  1. *InftyReade*
     http://ww
  2. *InftyReade*
  3. *BatchInfty*
  4. *CharImageManager*



"InftySystem" for large scale digitization

- Process Flow using *BatchInfty* & *InftyReader pro*

1. Noise reduction, centering, etc.
2. Trial recognition
3. Extraction features:
   - Document style → Logical structure analysis
   - Character cluster images → OCR engine
4. Recognition & verification
5. PDF output

Problems

- Full automatization of the adaptive method

From the target documents:
Extraction of character features / layout parameters

Without manual correction

Improvement of
- Character recognition
- Formulae recognition
- Logical structure analysis

Problems

- Further improvement of character/symbol recognition and structure analysis of math expressions.
  - Touched characters, Broken characters in math area
  - Low resolution image
  - Different type face (Old books, typewriter prints, etc.)
  - Bold char detection in math area

Problems

- Logical Structure Analysis (Automatic detection and manual correction)
  - Title, Autor, Section, Subsection, Itemization, BibItem, Theorem, Lemma, etc.
  - Hyperlink inside document.

Problems

- Detection/Analysis of Figures and Tables
  - Detection of characters in figures
  - Table structure analysis
  - Graphs → Tables

Challenge

- Is it possible to realize:

  OCR with higher accuracy
  than manual imput/correction by human?

Challenge

- Is it possible to realize:

  OCR with higher accuracy
  than manual imput/correction by human?

(I hope a student who challenges to this difficult problem appears in near future!)

Conclusion

- InftyProject.
  - Research group of math information processing.
- Demo (*InftyReader*).
- A Brief sketch of the methods used in Infty and the current state of the art.
  - There are many problems unsolved, especially in practical sense.
- Proposed some problems to be attacked.

Thanks you!

Masakazu Suzuki
Graduate School/ Faculty of Math.
Kyushu University
E-mail: suzuki@math.kyushu-u.ac.jp

InftyProject: http://www.inftyproject.org
Science Accessibility Net: http://www.sciaccess.net