# Adaptive method for the digitization of mathematical journals

IMU-WDML Workshop

June 2, 2012, Washington DC

Masakazu Suzuki
Kyushu University, Professor emeritus
Kyushu Institute of Systems, Informatics and Nanotechnologies (ISIT)
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
- Current State of the Art with Demonstration
- Adaptive Method
  - Character and Symbol Recognition
  - Logical Structure Analysis
- Future Problems

# *Section 1*

# About Infty Project

# InftyProject

- R&D on *Math Information Systems*
- Main system development

  InftyReader : Math OCR software

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

  ChattyInfty : InftyEditor + speech output, Authoring of DAISY

- URL :

  Project site: http://www.inftyproject.org/en//

  Release & user support of Infty products:
  Science Accessibility Net http://www.sciaccess.net/

# InftyProject

- R&D on *Math Information Systems*
- Main system development

  InftyReader : Math OCR software

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

  ChattyInfty : InftyEditor + speech output, Authoring of DAISY

- URL :

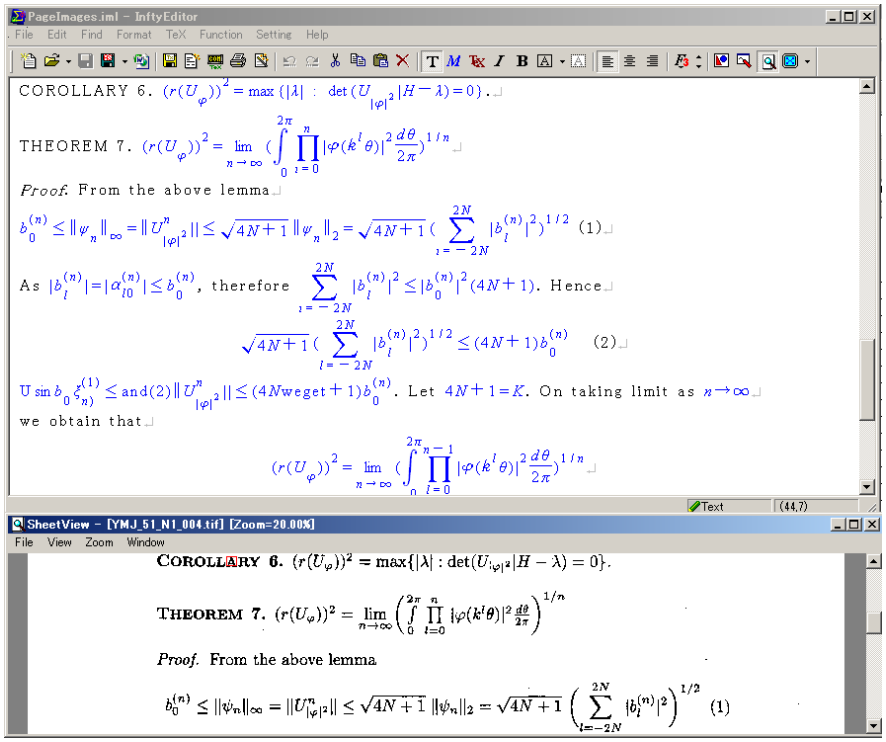
  Project site: http://www.inftyproject.org/en//

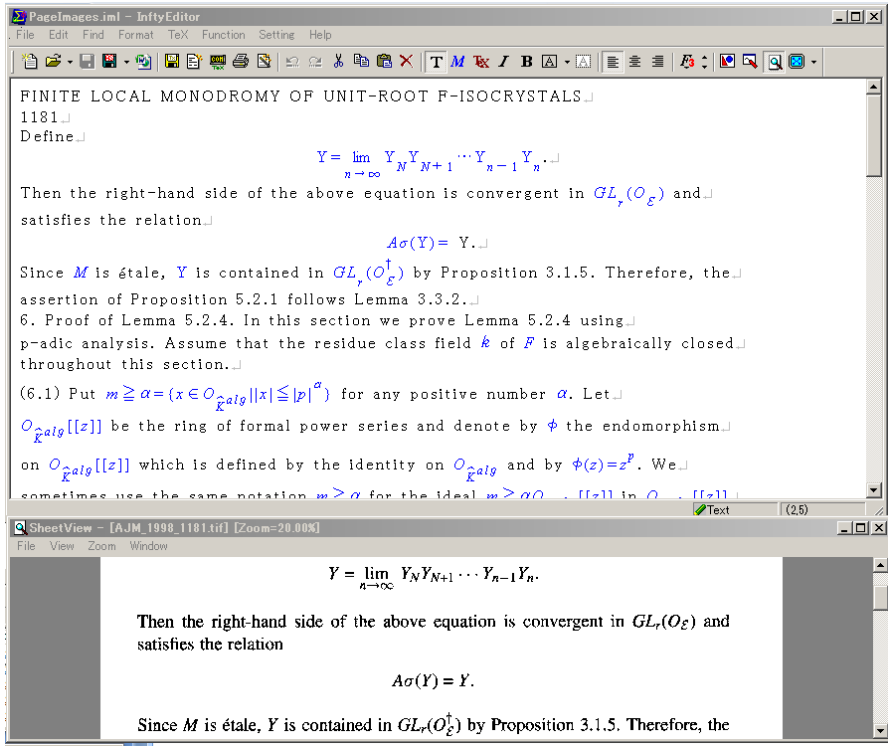
  Release & user support of Infty products:
  Science Accessibility Net http://www.sciaccess.net/

# "InftyReader" OCR software for math documents

- Demonstration.

  Recognition result samples (YMJ, AJM).



# *Section 2*

# Toward Rich DML

# 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, Bib Link
- Level 3: Structured accessible document
  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:
  e.g. GIF, T

WDML achieved this level.

- Level 2: Searchable digitized document
  e.g. PDF with hidden text, Bib Link
- Level 3: Structured accessible document
  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:
  e.g. PDF w

Infty : Level 1 → Level 3

- Level 3: Structured *accessible* document
  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

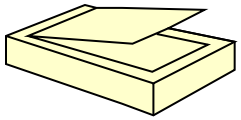


Image File (TIF)

PDF

Texts & Math symbols

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, XHTML+MathML,
PDF, Braille codes, etc.

# Layout Analysis

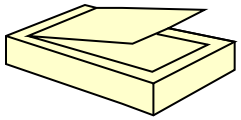
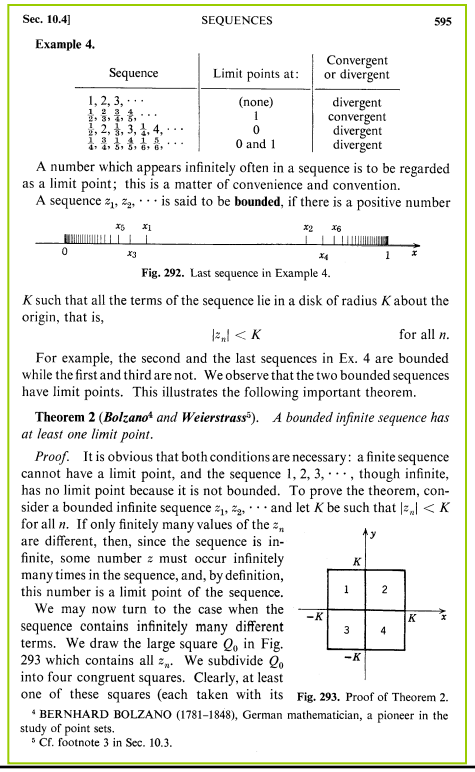


Image File (TIF)

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 |
| $\frac{1}{2}, 2, \frac{3}{4}, 3, \frac{5}{6}, 4, \cdots$ | 0 | divergent |
| $\frac{1}{2}, \frac{2}{3}, \frac{3}{4}, \frac{4}{5}, \frac{5}{6}, \frac{6}{7}, \cdots$ | 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.

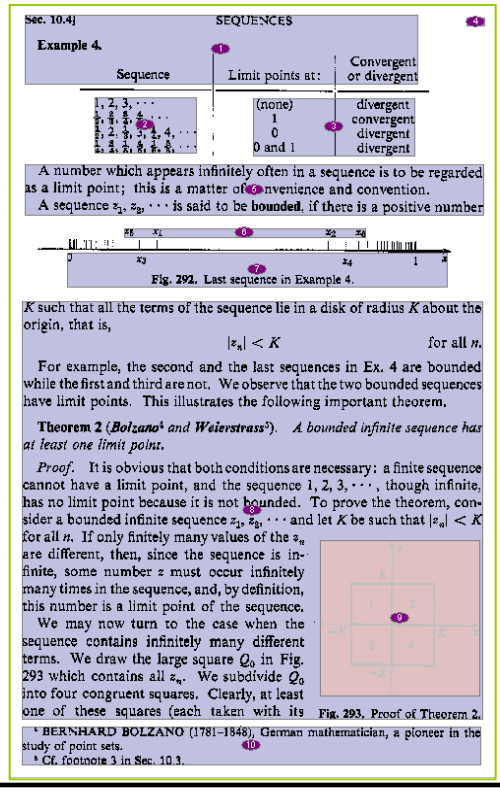
We may now turn to the case when the sequence contains infinitely many different terms. We draw the large square $Q_0$ in Fig. 293 which contains all $z_n$. We subdivide $Q_0$ into four congruent squares. Clearly, at least one of these squares (each taken with its




**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.



Sec. 10.4] SEQUENCES 4

**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 |
| $\frac{1}{2}, 2, \frac{3}{4}, 3, \frac{5}{6}, 4, \cdots$ | 0 | divergent |
| $\frac{1}{2}, \frac{2}{3}, \frac{3}{4}, \frac{4}{5}, \frac{5}{6}, \frac{6}{7}, \cdots$ | 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

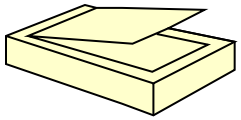


Image File (TIF)

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

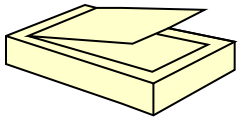


Image File (TIF)

PDF

Texts & Math symbols

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, XHTML+MathML,
PDF, Braille codes, etc.

# Process Flow of Digitization

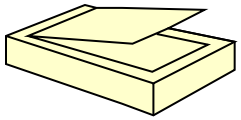


Image File (TIF)

PDF

Texts & Math symbols

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, XHTML+MathML,
PDF, Braille codes, etc.

# Process Flow of Digitization

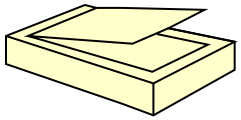


Image File (TIF)

PDF

Texts & Math symbols

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, XHTML+MathML,
PDF, Braille codes, etc.

# Document Structure Analysis

- Detection of :

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

  - Currently, naïve methods are used:

    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.

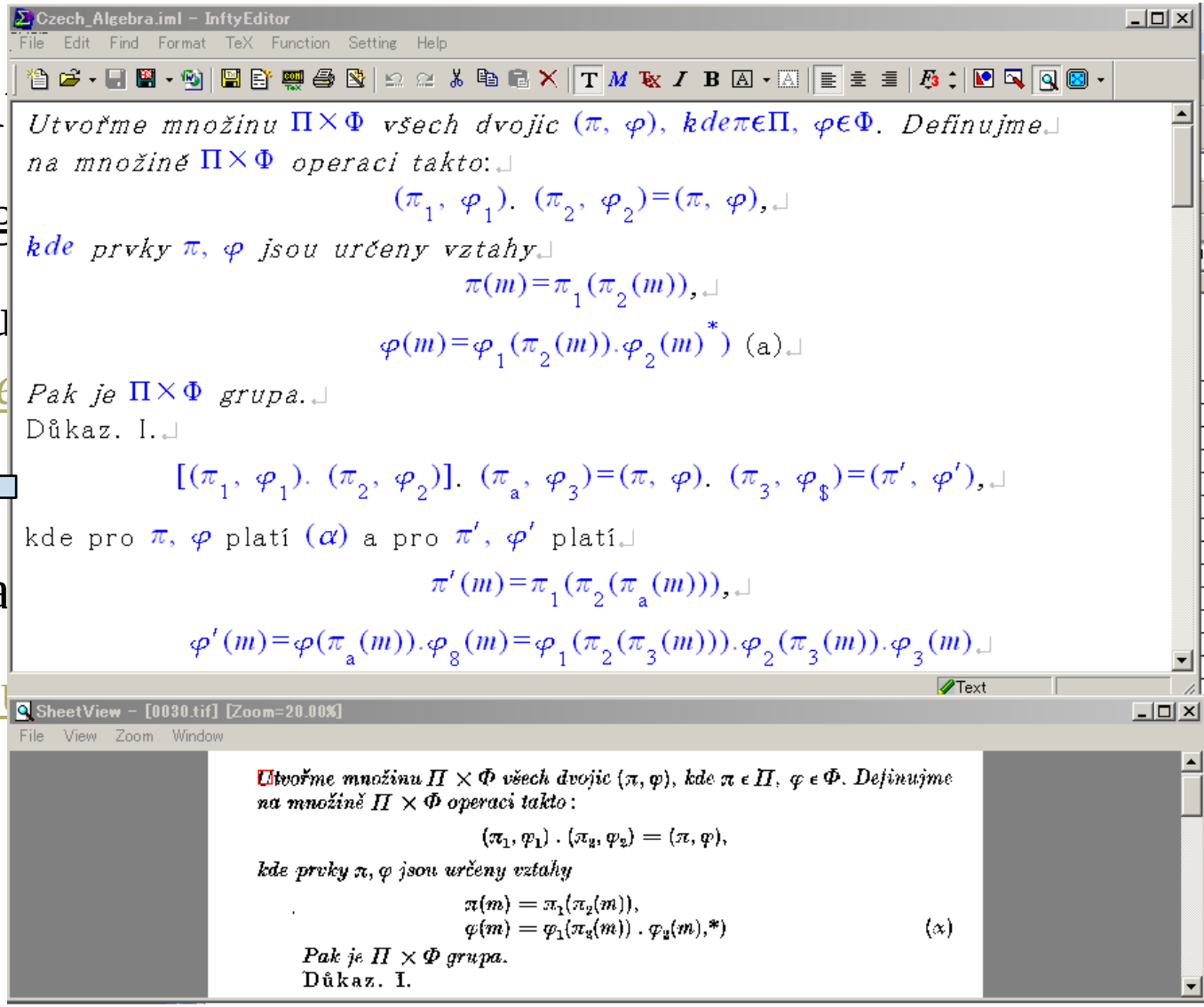
# *Section 3*

# Current state of the art with demonstration

- Demonstration…
  - Math recognition *(Already shown)*
  - Multi lingual recognition ← *FineReader OCR plug-in*
    *Czech paper result sample*
  - Matrices
  - Layout analysis, Table recognition
  - Logical structure analysis

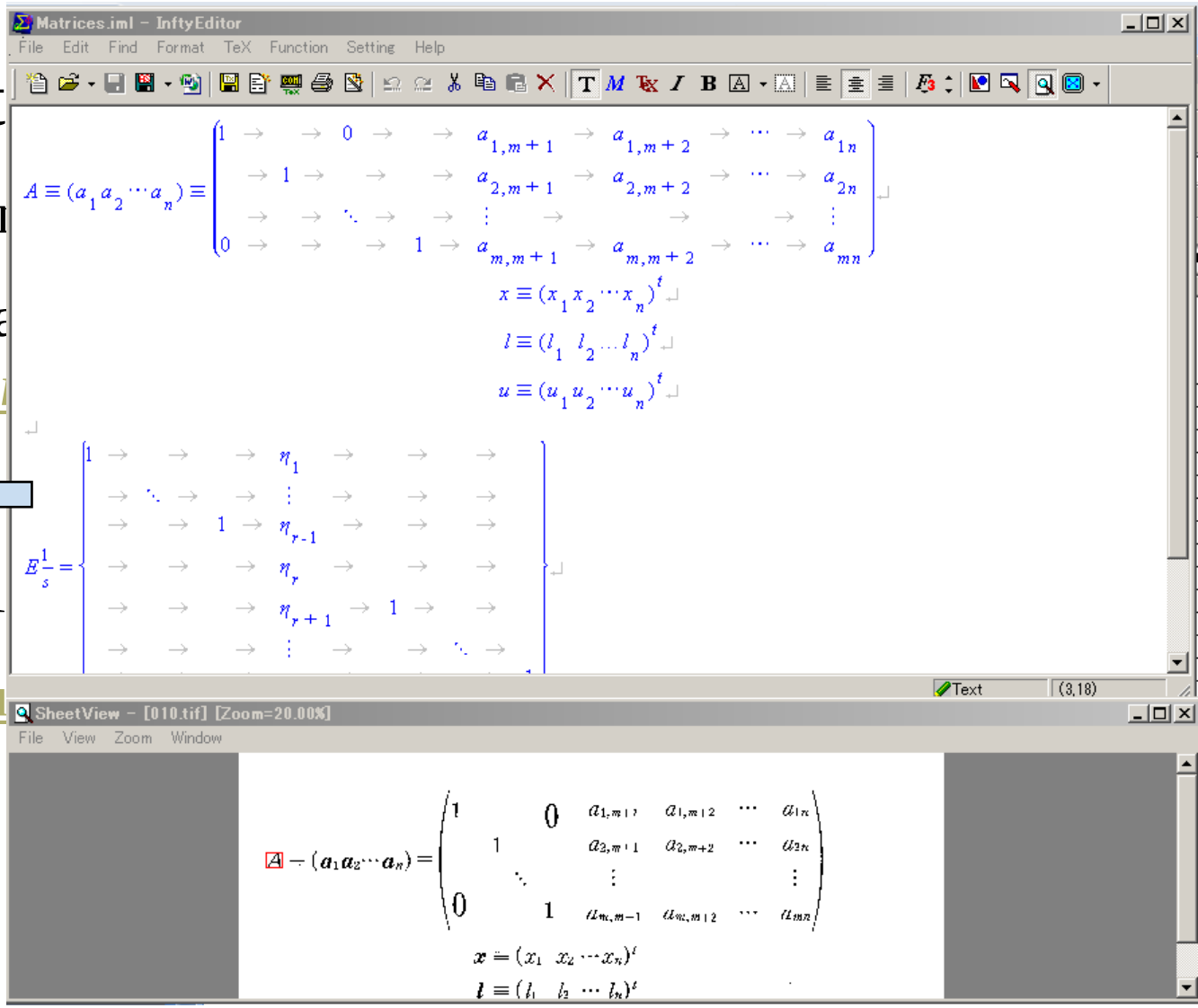
# "InftyReader" OCR software for math documents

- Demonstra
  - Math recog
  - Multi lingu
    Czech pape
  - Matrices
  - Layout ana
  - Logical str



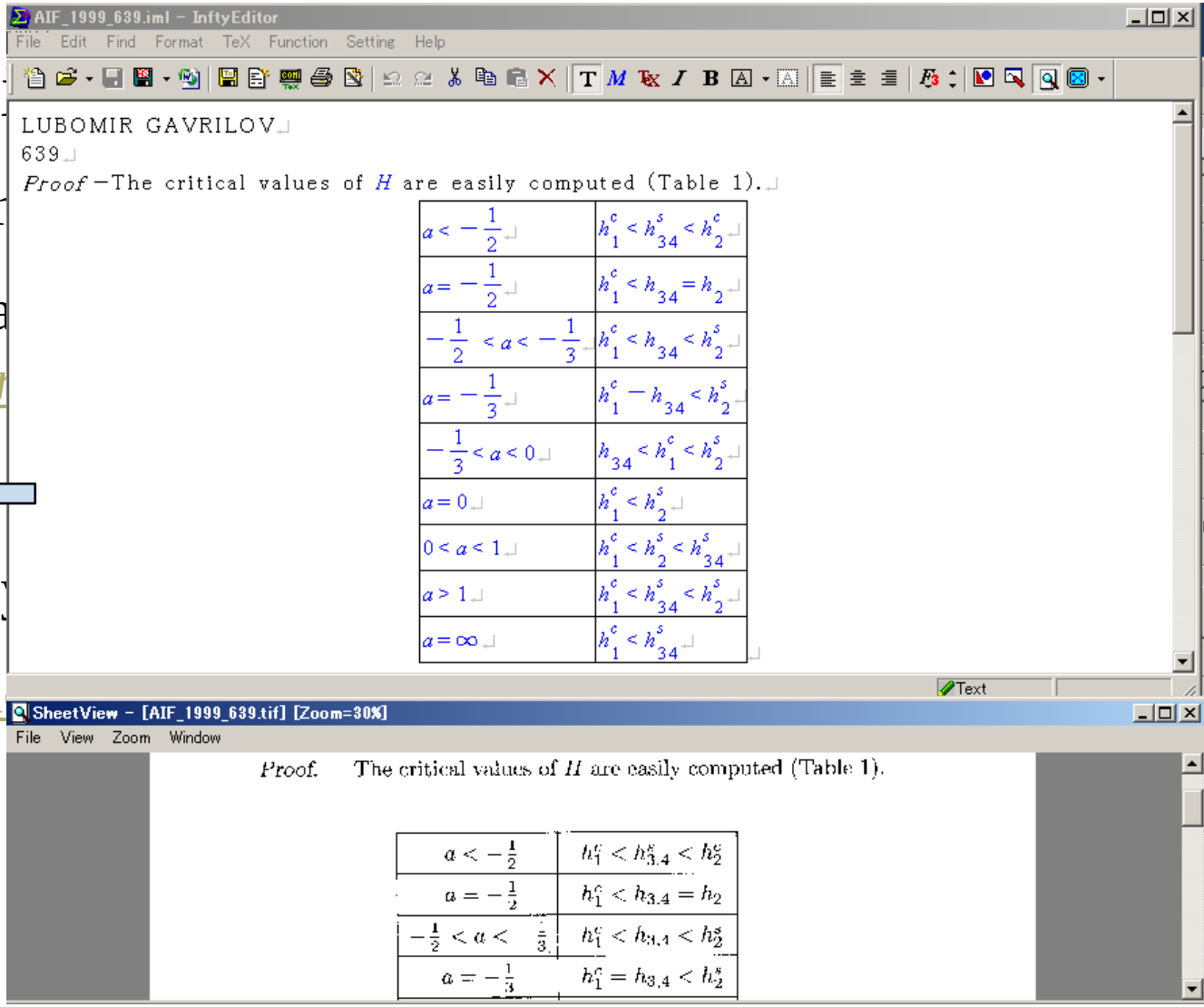
# “InftyReader” OCR software for math documents

- Demonstrat
  - Math recogn
  - Multi lingua
    *Czech papei*
  - Matrices
  - Layout anal
  - Logical stru



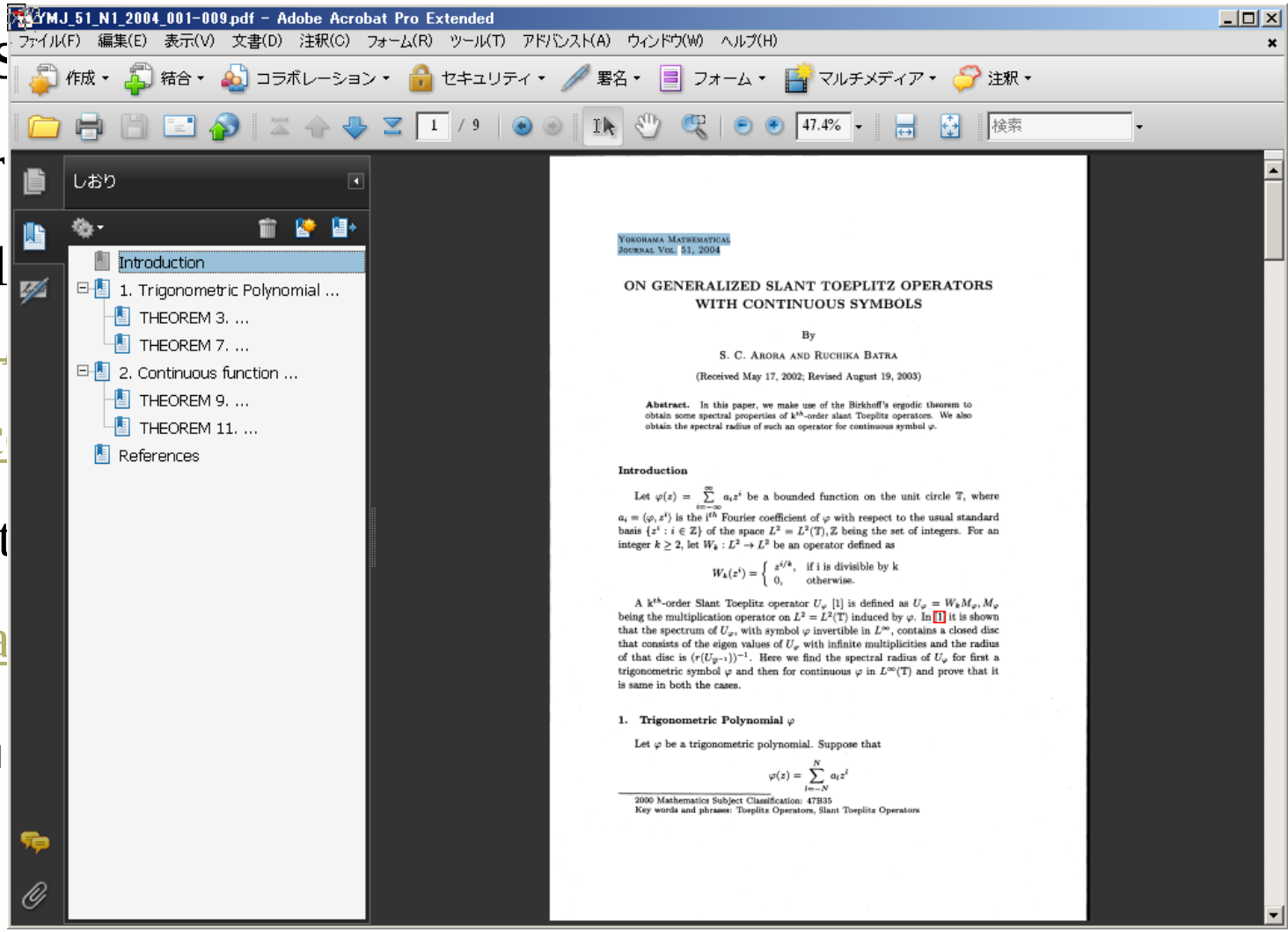
# "InftyReader" OCR software for math documents

- Demonstrat
  - Math recogn
  - Multi lingua
    *Czech paper*
  - Matrices
  - Layout anal
  - Logical stru



# "InftyReader" OCR software for math documents

- Demons
  - Math r
  - Multi l
    *Czech*
  - Matric
  - Layout
  - Logica



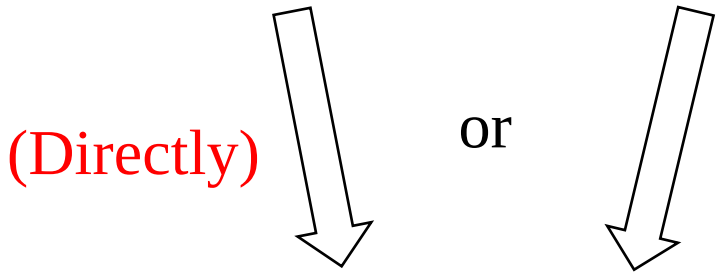
- Demonstration…
  - Math recognition *(Already shown)*
  - Multi lingual recognition ← *FineReader OCR plug-in*
    *Czech paper result sample*
  - Matrices
  - Layout analysis, Table recognition
  - Logical structure analysis

# *Section 4*
# Large Volume Recognition

# Large Volume Digitization

- *Adaptive method* is efficient:

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




Recognition

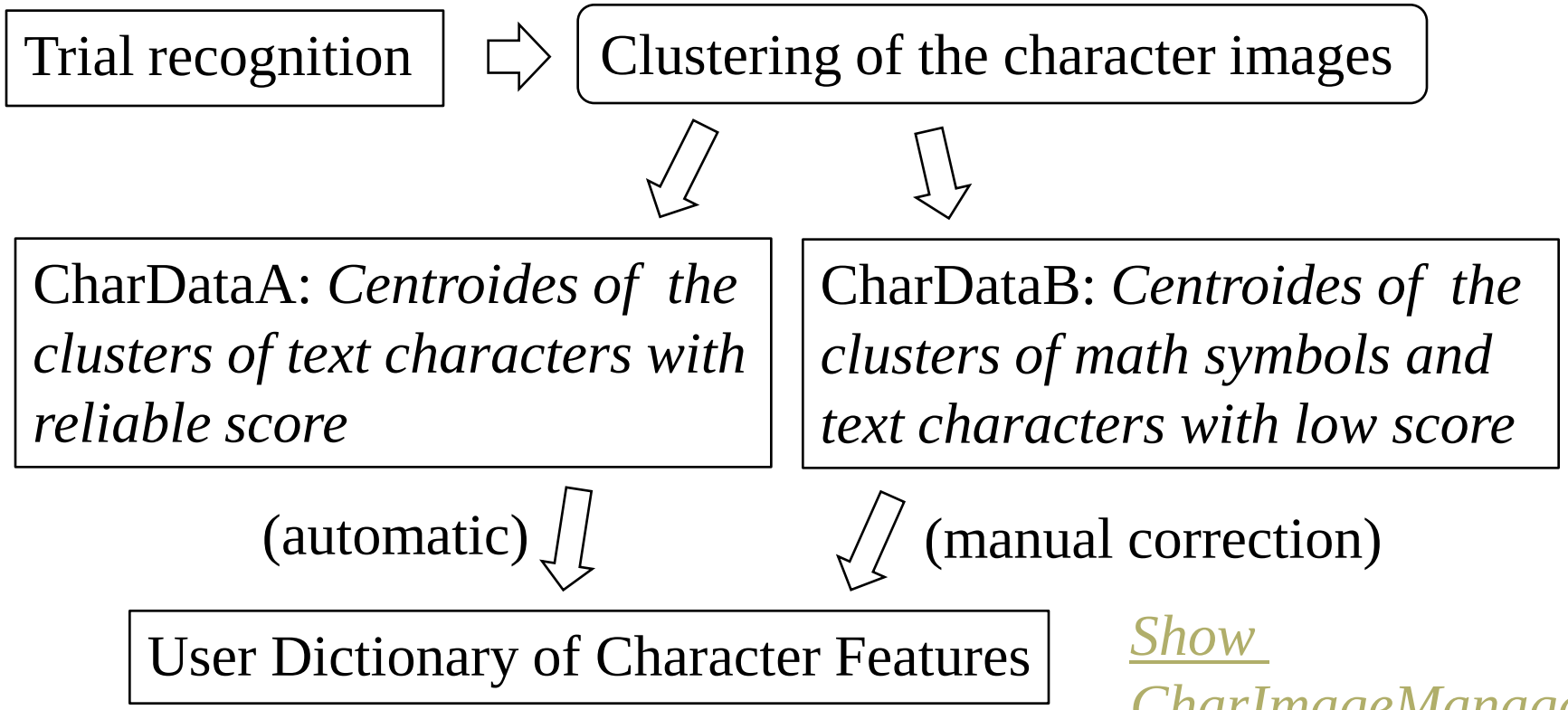
# Large Volume Digitization

- 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

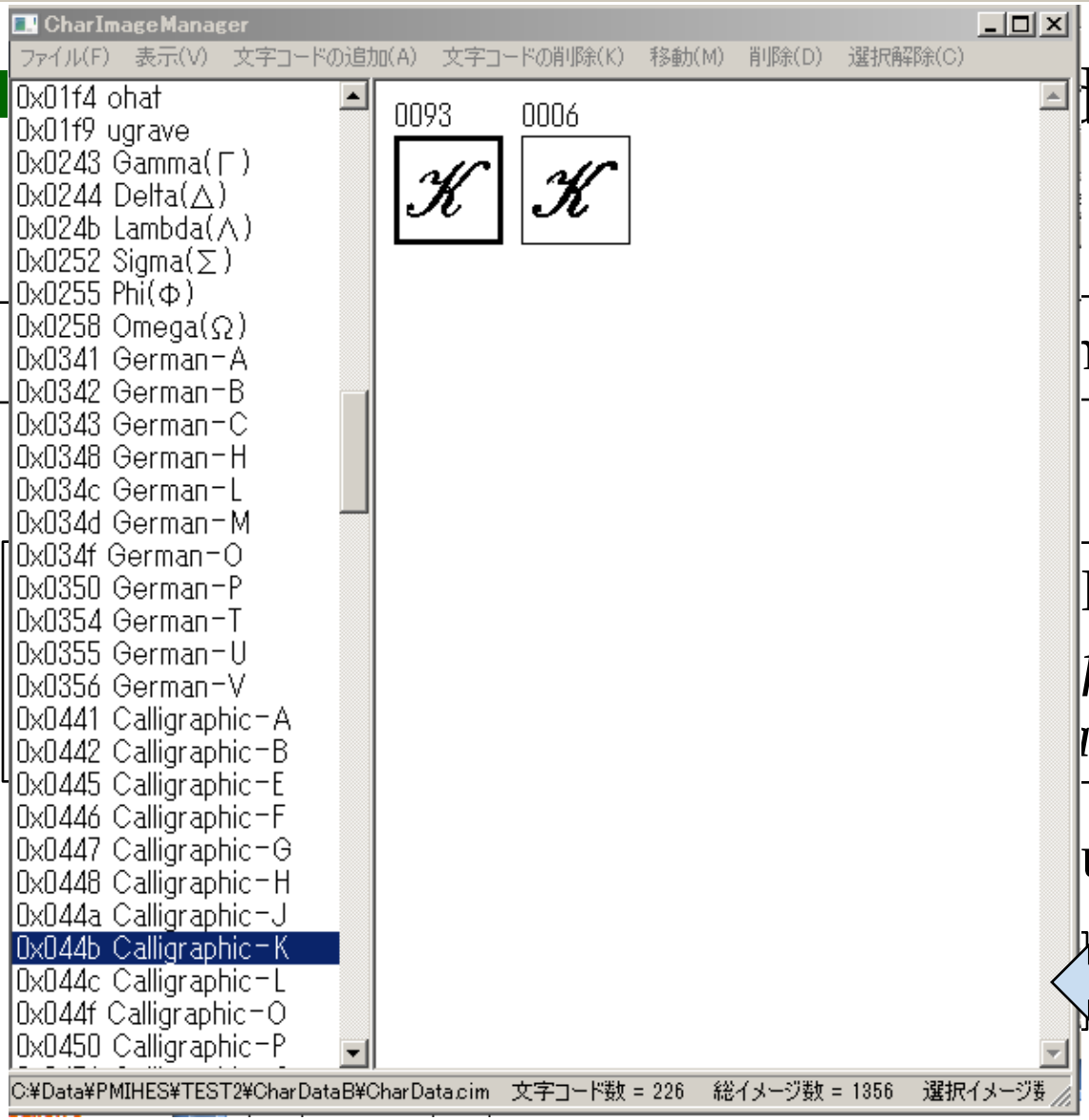
# Large Volume Digitization

- Generation of *UserDictionary* adapting OCR engine to the target documents.




*Show CharImageManager*

# Large Volume Digitization




lapting OCR

haracter images

B: *Centroides of the*
*f math symbols and*
*icters with low score*

ual correction)

*Show CharImageManager*

# *Section 5*
# Open Problems

# 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

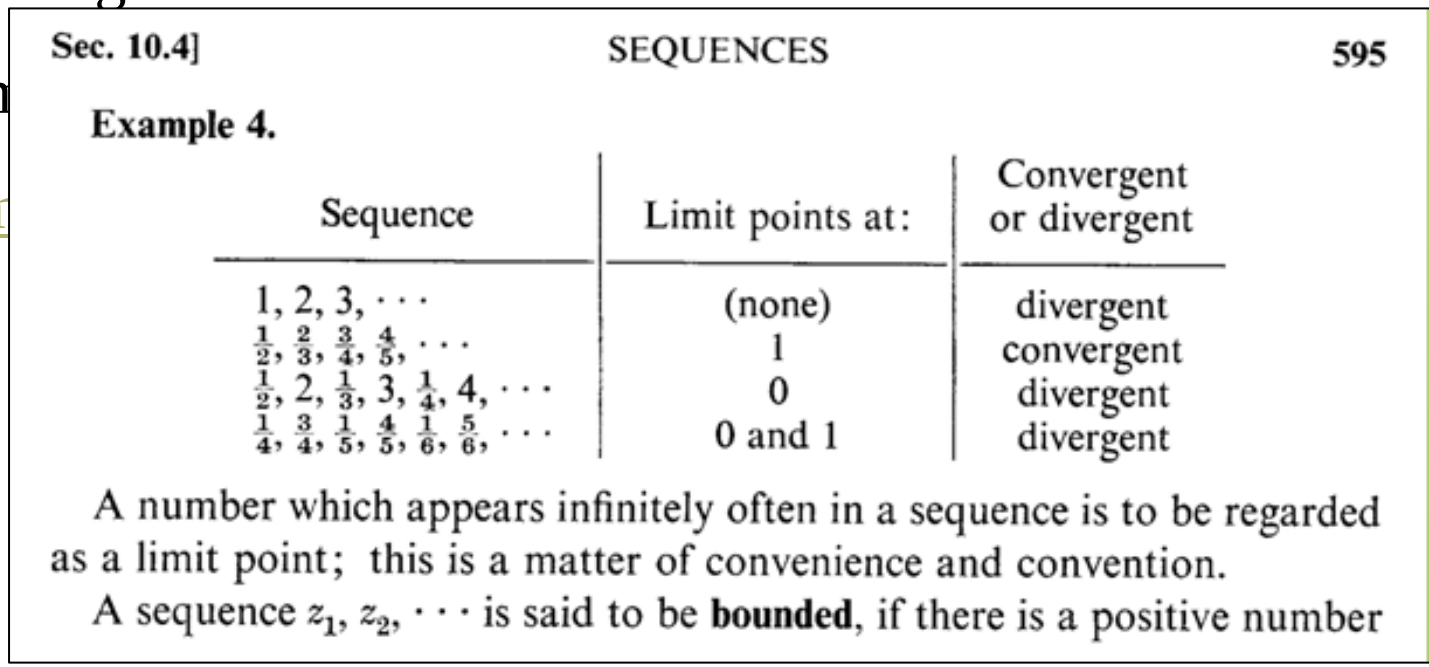
- Logical Structure Analysis (Automatic detection and manual correction) --- *still difficult!*
  - Title, Autor, Section, Subsection, Itemization, BibItem, Theorem, Lemma, etc.
  - Hyperlink inside document.

# Problems

- Detection/Analysis of Figures and Tables
  - Detection of characters in figures
  - Table structure analysis *(Sample)*
  - Diagram recognition
    - Chemical diagrams ← *Recently developing world wide*
    - (Commutative diagrams) ← *Future work*

# Problems

- Detection/Analysis of Figures and Tables
  - Detection of characters in figures
  - Table structure analysis (*Sample*)
  - Diagram recognition
    - Chen
    - (Com



Sec. 10.4] SEQUENCES 595

**Example 4.**

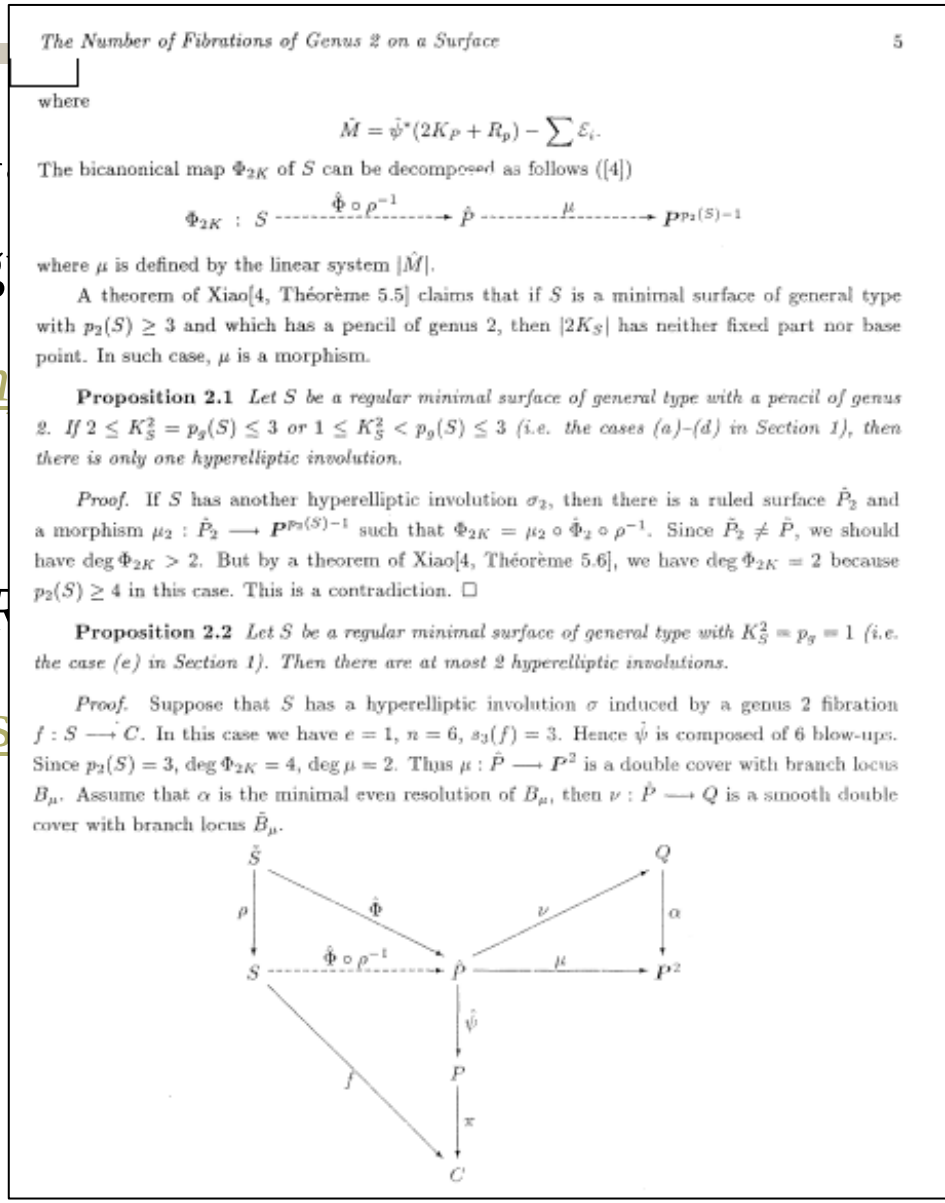
| Sequence | Limit points at: | Convergent or divergent |
|---|---|---|
| $1, 2, 3, \cdots$ | (none) | divergent |
| $\frac{1}{2}, \frac{2}{3}, \frac{3}{4}, \frac{4}{5}, \cdots$ | 1 | convergent |
| $\frac{1}{2}, 2, \frac{1}{3}, 3, \frac{1}{4}, 4, \cdots$ | 0 | divergent |
| $\frac{1}{4}, \frac{3}{4},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

# Problems

- Detection/Analysis of Fi
  - Detection of characters in fig
  - Table structure analysis (*Sam*
  - Diagram recognition
    - Chemical diagrams ← *K*
    - (Commutative diagrams



*The Number of Fibrations of Genus 2 on a Surface* 5

where

$$\hat{M} = \hat{\psi}^*(2K_P + R_p) - \sum \mathcal{E}_i.$$

The bicanonical map $\Phi_{2K}$ of $S$ can be decomposed as follows ([4])

$$\Phi_{2K} \ : \ S \xrightarrow{\hat{\Phi} \circ \rho^{-1}} \hat{P} \xrightarrow{\mu} P^{p_2(S)-1}$$

where $\mu$ is defined by the linear system $|\hat{M}|$.

A theorem of Xiao[4, Théorème 5.5] claims that if $S$ is a minimal surface of general type with $p_2(S) \geq 3$ and which has a pencil of genus 2, then $|2K_S|$ has neither fixed part nor base point. In such case, $\mu$ is a morphism.

**Proposition 2.1** *Let $S$ be a regular minimal surface of general type with a pencil of genus 2. If $2 \leq K_S^2 = p_g(S) \leq 3$ or $1 \leq K_S^2 < p_g(S) \leq 3$ (i.e. the cases (a)–(d) in Section 1), then there is only one hyperelliptic involution.*

*Proof.* If $S$ has another hyperelliptic involution $\sigma_2$, then there is a ruled surface $\hat{P}_2$ and a morphism $\mu_2 : \hat{P}_2 \longrightarrow P^{p_2(S)-1}$ such that $\Phi_{2K} = \mu_2 \circ \hat{\Phi}_2 \circ \rho^{-1}$. Since $\hat{P}_2 \neq \hat{P}$, we should have $\deg \Phi_{2K} > 2$. But by a theorem of Xiao[4, Théorème 5.6], we have $\deg \Phi_{2K} = 2$ because $p_2(S) \geq 4$ in this case. This is a contradiction. □

**Proposition 2.2** *Let $S$ be a regular minimal surface of general type with $K_S^2 = p_g = 1$ (i.e. the case (e) in Section 1). Then there are at most 2 hyperelliptic involutions.*

*Proof.* Suppose that $S$ has a hyperelliptic involution $\sigma$ induced by a genus 2 fibration $f : S \longrightarrow C$. In this case we have $e = 1$, $n = 6$, $s_3(f) = 3$. Hence $\hat{\psi}$ is composed of 6 blow-ups. Since $p_2(S) = 3$, $\deg \Phi_{2K} = 4$, $\deg \mu = 2$. Thus $\mu : \hat{P} \longrightarrow P^2$ is a double cover with branch locus $B_\mu$. Assume that $\alpha$ is the minimal even resolution of $B_\mu$, then $\nu : \hat{P} \longrightarrow Q$ is a smooth double cover with branch locus $\bar{B}_\mu$.

# Conclusion

- InftyProject.
  - Research group of math information processing.
- Demo (*InftyReader*) to show the current state of the art.
- Adaptive method to improve character and symbol recogition *(CharImageManager)*.
- Proposed some problems to be attacked.

# Thanks you!

Masakazu Suzuki
suzuki@isit.or.jp (current address)
msuzuki@kyudai.jp (permanent address)

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