Adaptive Method for the Digitization of Mathematical Journals

September 9, 2009

Kyoto University Library

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

Plan of the talk

- About InftyProject and sAccessNet
- Digitization of Mathematical Journals
  - Different levels of digitization
  - Process Flow of Digitization
  - Adaptive Method
  - Demo

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

- Non profit organization “Science Accessibility Net”
  - Helping people with visual handicaps working in scientific fields.
  - Digitization of mathematical/scientific documents (Journals and books)
- http;//www.sciaccess.net/

Go

Digitization of Math Journals

- Search

  Bibliographic data, Text, Structure, 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

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 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

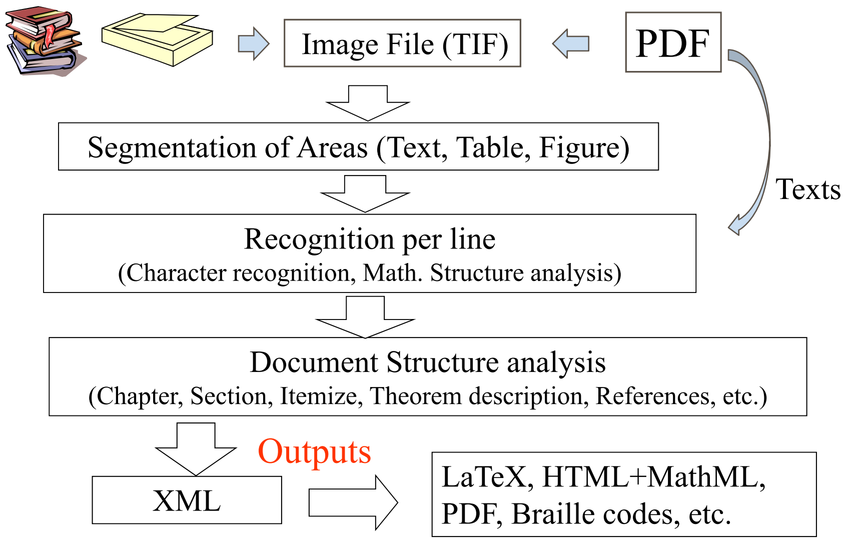
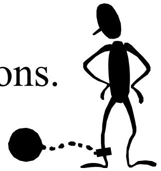


Image File (TIF)
PDF
Segmentation of Areas (Text, Table, Figure)
Texts
Recognition per line
(Character recognition, Math. Structure analysis)
Document Structure analysis
(Chapter, Section, Itemize, Theorem description, References, etc.)
Outputs
XML
LaTeX, HTML+MathML,
PDF, Braille codes, etc.

Difficulty of Math. recognition

- Symbols (Greeks, various math. symbols…)
- Fotns (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



“INFTY” *an integrated OCR for mathematical documents*

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

“INFTY” *an integrated OCR for mathematical documents*

- 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

“INFTY” *an integrated OCR for mathematical documents*

- 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
- Demonstration …

おわり

Thanks you!

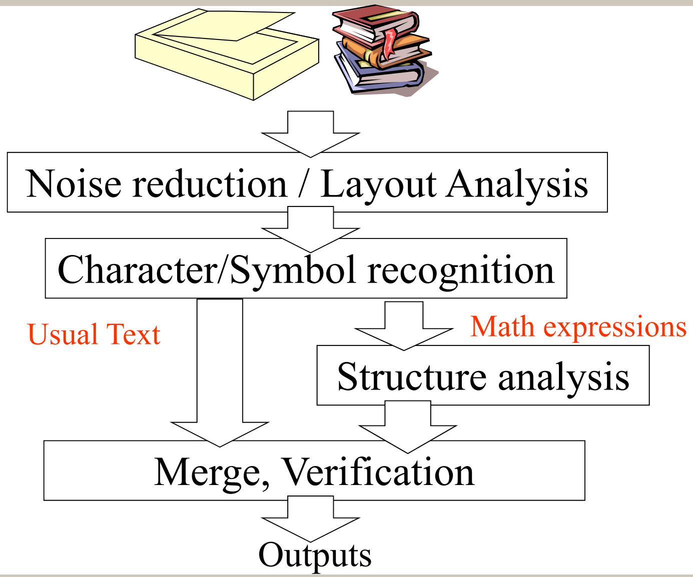
InftyProject: http://www.inftyproject.org/
sAccessNet: http://www.sciaccess.net/

Difficulty of Math. recognition

- Symbols (Greeks, various math. symbols…)
- Fotns (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. ex
- Distinction of noises and small sy

Adaptive method is Efficient！

INFTY’s Recognition Flow



Noise reduction / Layout Analysis

Character/Symbol recognition

Usual Text

Math expressions

Structure analysis

Merge, Verification

Outputs

“InftyReaderPro” *Digitization of math journals*

- User interface to edit recognition results *keeping the coordinates* of the characters in the original images,

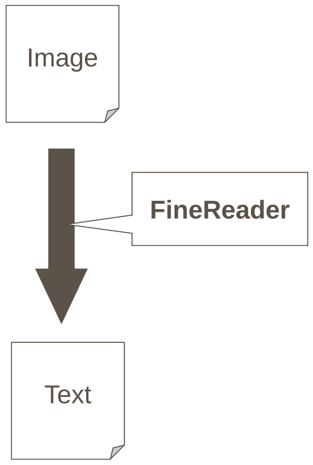
“InftyReaderPro” *Digitization of math journals*

- User interface to edit recognition results *keeping the coordinates* of the characters in the original images, and
- To edit logical structures and hyperlinks.

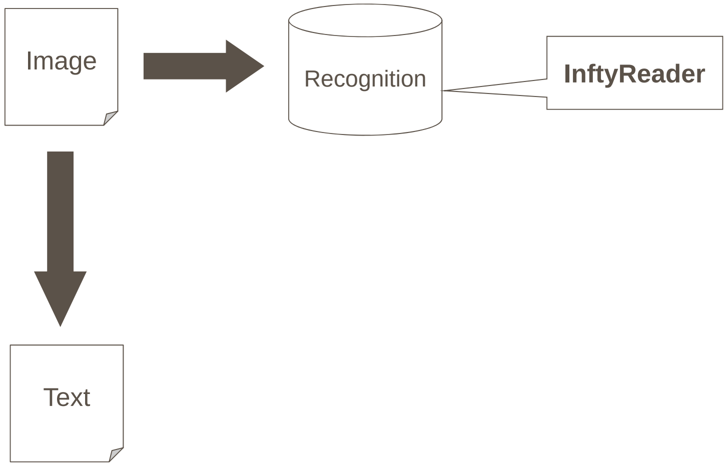
“BatchInfty” *Digitization of math journals*

- Processing large volumes of journals
- Integration of any other OCR (e.g. FineReader) into InftyReader

“BatchInfty” *Digitization of math journals*



“BatchInfty” *Digitization of math journals*



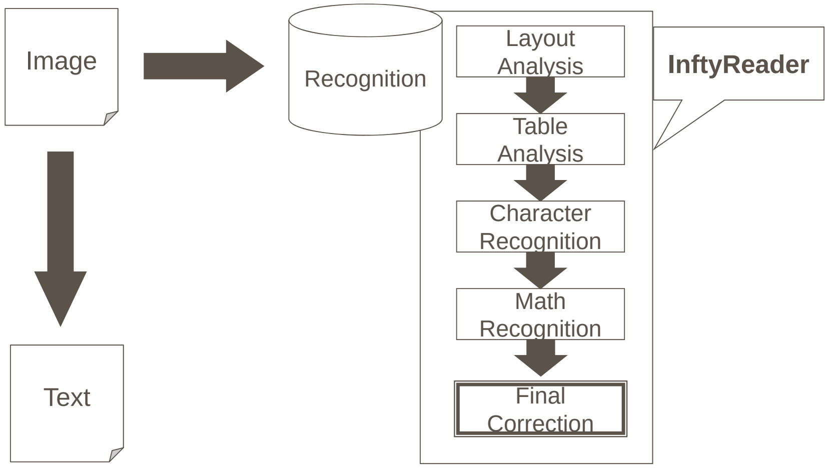
Image

Recognition

InftyReader

Text

“BatchInfty” *Digitization of math journals*



Image

Recognition

Layout Analysis

Table Analysis

Character Recognition

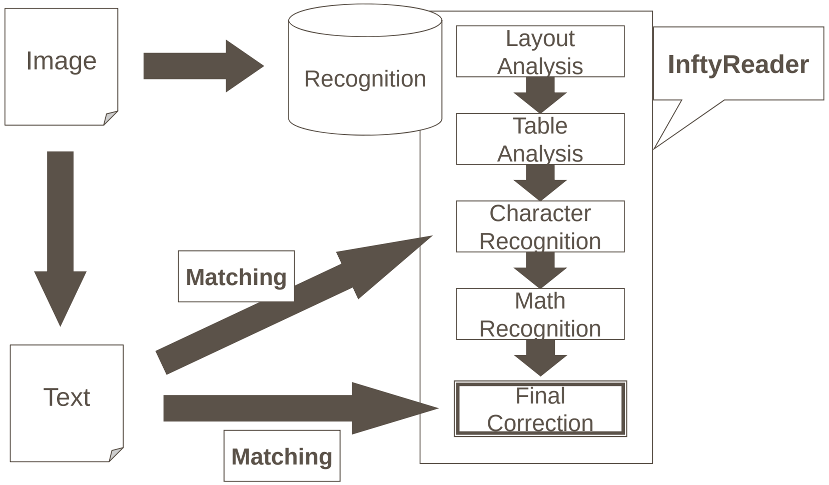
Math Recognition

Final Correction

InftyReader

Text

“BatchInfty” *Digitization of math journals*



Image

Recognition

Layout Analysis

Table Analysis

Character Recognition

Math Recognition

Final Correction

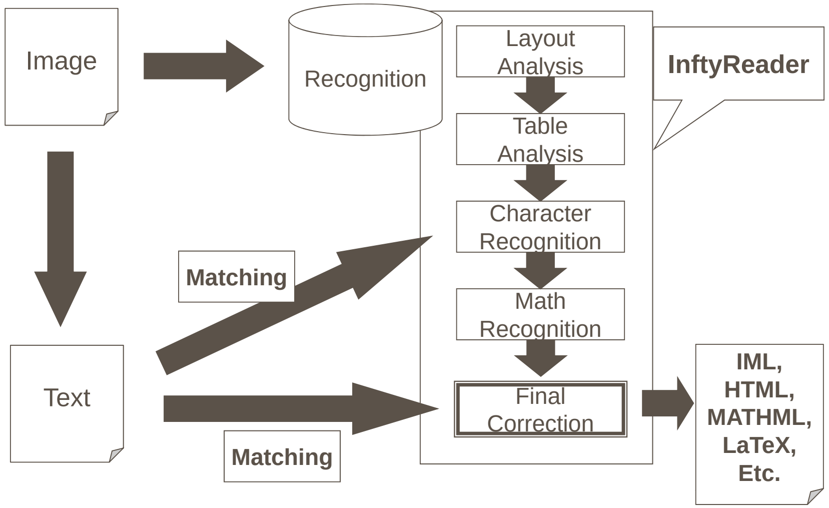
InftyReader

Matching

Text

Matching

“BatchInfty” *Digitization of math journals*



Image

Recognition

Layout Analysis

Table Analysis

Character Recognition

Math Recognition

Final Correction

InftyReader

Matching

Text

Matching

IML, HTML, MATHML, LaTeX, Etc.

“BatchInfty” *Digitization of math journals*

- Processing large volumes of journals
- Integration of any other OCR (e.g. FineReader) into InftyReader

“BatchInfty” *Digitization of math journals*

- Processing large volumes of journals
- Integration of any other OCR (e.g. FineReader) into InftyReader
- Extract logical structures of each articles

  → Table of contents, Hyper links.

BatchInfty & InftyReader Pro

- Demonstration …