Integrated OCR software for mathematical documents and its output with accessibility

Masakazu Suzuki[1], Toshihiro Kanahori[2], Nobuyuki Ohtake[2], and Katsuhito Yamaguchi[3]

[1] Kyushu University, Faculty of Mathematics,
Hakozaki 6-10-1, Higashiku, Fukuoka, 812-8581 Japan
suzuki@math.kyushu-u.ac.jp
http://www.math.kyushu-u.ac.jp/ suzuki/index.html
[2] Research Center on Educational Media, Tsukuba College of Technology,
4-12 Kasuga, Tsukuba-shi, Ibaraki, 305-0821 Japan
{kanaori,ohtake}@k.tsukuba-tech.ac.jp
[3] Nihon University, Junior College Funabashi Campus,
7-24-1, Narashinodai, Funabashi, Chiba 274-8501, Japan
eugene@gaea.jcn.nihon-u.ac.jp

Abstract. This paper describes shortly a practical integrated system for scientific documents including mathematical formulae, named 'Infty'. The system consists of three components of applications: an OCR system named 'InftyReader', an editor named 'InftyEditor' and converting tools into various formats. Those applications are linked each other via XML files.

InftyReader recognizes scanned images of clearly printed mathematical documents and outputs the recognition results in a XML format. It recognizes complex mathematical formulae used in various research papers of mathematics including matrices. InftyEditor provides a very efficient interface to correct the recognition results using keyboard. Another feature of InftyEditor is its handwriting interface to input mathematical formulae for users with vision and speech interface for visually disabled uses.

The XML files output by InftyReader/Editor can be converted into various formats: LaTeX, MathML, HTML and Braille Codes; in UBC (Unified Braille Codes) for English texts and in Japanese Braille Codes for Japanese texts.

1 Introduction

We presented a prototype of mathematical document recognition system in [1] at ICCHP 2000. As a result of the improvements of recognition rates and user interfaces obtained during this period, we released recently the software 'InftyReader' to transcribe printed mathematical documents into digital data with accessibility as freely usable software for non-profit purpose. The purpose of this paper is to describe the outline of the software[1]. A package file of the software can be obtained from the WEB site [7].

[1] In the lecture, there will be a demonstration of the system using real data.

InftyReader recognizes clearly printed papers, carefully scanned into binary images by either 600 DPI or 400 DPI. The image files have to be prepared in TIFF CCITT-3 or CCITT-4 format. InftyReader segments page images into picture areas, table areas and text areas, and then recognizes text areas including mathematical expressions. To get better recognition results, users are recommended to adjust the binarization threshold of their scanner so that, in scanned page images, the number of the touched or broken characters be less than 1% of the total number of the characters in each page.

The released version of the software includes an editor of mathematical documents to view and edit the recognition results, named "InftyEditor". InftyEditor is typesetting tool of scientific documents having, in addition to usual keyboard interface by TEX-like command input, a handwriting interface to input mathematical expressions. Users can edit easily the recognition results, and save the results as a LATEXsource file, a HTML file, or in a text file using a "Human Readable TEX" format for visually disabled persons. The trial version of InftyReader/Editor provides also Braille outputs as well as a speech interface to read and edit mathematical documents for visually disabled users. The description of the software below is based on the InftyReader Version 2.4.2 released on April 15, 2004.

The **Fig. 1** below shows a snapshot of a recognition result displayed on the InftyEditor incorporated in InftyReader. Since the Infty system consists of three components of applications: an OCR system named 'InftyReader', an editor named 'InftyEditor' and a converter unit of Math-XML into various formats, we shall describe the system in three steps below.

2 Optical Recognition of Printed Mathematical Documents

In this section, we shall describe briefly some specification of our software 'InftyReader' to recognize mathematical documents. As for the details about the recognition methods used and experimental results, see [2], [3] and [4]. Especially, the paper [4] reports a detailed experimental results of InftyReader on 496 page images taken from 25 different volumes of pure mathematical journals and a book from physics, including various levels of scanning quality from good ones to very noisy pages having more than 10 percent abnormal (touched/broken) characters.

One of the important result of the experiments in [4] and the experiments by the current version of InftyReader is the strong relationship between the total recognition rates of character recognition and the ratio of the number of abnormal (broken/touched) characters included mathematical formulae. There was a remarkable difference on the recognition rates between the scanned images with abnormal (broken/touched) characters less than 1% of the total number of characters in each page and the noisy pages images containing more abnormal (broken/touched) characters by the test using our database. On the other hand, in our experience, usually it is not difficult to reduce the number of abnormal

characters less than 1% by careful scanning, if the target document is clearly printed. Almost a half of the articles in the test data mentioned above are of this quality.

Note that there is an essential difference between the current version of InftyReader and the version used in the experiments reported in [4]. While the version of InftyReader reported in [4] used a commercial OCR engine for both Japanese characters and alpha-numeric characters in usual text area, the current version of InftyReader makes use of the commercial OCR engine to recognize only Japanese characters[2]. Its English version makes use of no commercial OCR engine and uses originally developed OCR to recognizes alpha-numeric characters and various mathematical symbols. However, as far as in our experiments performed on the same data mentioned above, the total recognition rates of the new version is even better than the old one, by the implementation of a touched character separation method valid both in text area and mathematical formulae areas. The average recognition rate of the new version for 13 articles with abnormal (broken/touched) characters less than 1% of the total number of characters in the database mentioned above was 99.89% for ordinary text areas and 99.56% including characters and symbols in mathematical formulae.

From each page image obtained by a scanner, InftyReader first cuts out picture areas, table areas and text areas automatically. Mathematical formulae are included in text areas in the first segmentation. As for the tables and the pictures, InftyReader outputs only brief information concerning their positions and their sizes.

For math-text area, InftyReader separates the area into lines and each line into ordinary text parts and the mathematical formulae parts using the recognition results of characters together with their positions and their sizes. In the case of English documents, linguistic information such as fundamental English words dictionary with short spelling, mathematical technical terms dictionary and some morphological analysis are used in order to improve the accuracy of the recognition of words and the separation of ordinary text parts and mathematical formulae parts.

InftyReader distinguishes more than 500 categories of characters and symbols, where Roman fonts of Upright, Italic, Caligraphic, Blackboard Bold are classified into different categories. However, it does not distinguish bold fonts with normal stroke fonts at all and the recognition rate of German fraktur or Script Roman fonts still remains on a very low level at the current stage of the development.

As for the recognition of mathematical formulae structure, we introduced a new algorithm to improve the robustness of the recognition against the variation of the styles of printing of the documents, using the spanning tree of minimum cost in the network generated by virtual links of characters [2]. InftyReader some times succeeds to recognize very complex matrices by the algorithm of [3] implemented in our system (see **Fig. 3** below at the end of the paper).

[2] We would like to express our thanks to Toshiba Coorporation for the supply of their OCR engine.

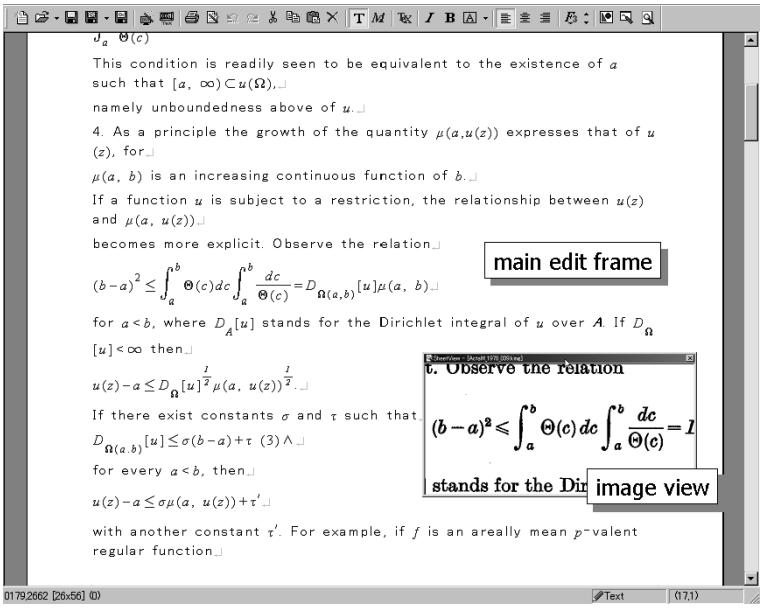

Fig. 1. Snapshot of InftyReader

3 Editor for Mathematical Documents – InftyEditor

The recognition results of InftyReader are output in a XML object and displayed in the main frame of the editor 'InftyEditor' included in the InftyReader package.

3.1 Interface for users with sight

InftyEditor is typesetting tool of scientific documents having a handwriting interface to input mathematical expressions, in addition to usual input interface by using palette or keyboard interface by TEX-like command input. Special edition of InftyEditor incorporated into InftyReader has a special view window to show the original scanned image to compare the recognition results with it (**Fig. 1**).

The displayed texts and mathematical formulae can be edited freely by using ordinary editing operations; cut, copy, paste and delete, and for mathematical symbols, our original front-end processor by LATEX commands is provided. For example, a fraction '$\frac{1}{x^2+1}$' can be input using only a keyboard in the following way;

1. input '\`frac`', then a fractional line appears and the cursor of InftyEditor moves to the numerator position of the fractional line,
2. input '1', and press the enter key, then the cursor moves to the denominator position,

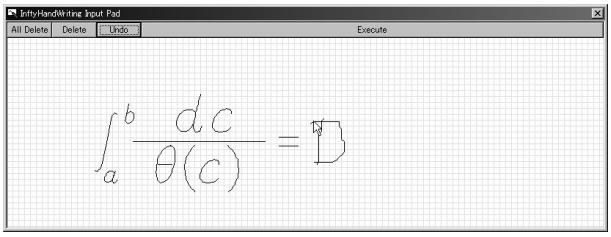

Fig. 2. Handwriting Dialog

3. input '**x**' and '^', then the cursor moves to the superscript position of 'x',
4. input '2', then the cursor moves to the right hand side position of 'x^2',
5. input '+' and '1', and press the enter key, then the cursor moves to the right hand side position of '$\frac{1}{x^2+1}$', and the input is done.

Another distinctive feature of InftyEditor is that it is provided with a handwriting interface to input mathematical formulae (**Fig. 2**). By clicking a button on the Toolbar, the dialog to input a pen stroke using a mouse (or data tablet / pen display) opens. The mathematical formulae written by hand in this dialog box is recognized and put into the mathematical text displayed on the main board of InftyEditor, at its cursor position. By this handwriting interface, the system realizes a very easy intuitive methods to edit or correct mathematical formulae, requiring no special skill about, for example, LaTeX notations of mathematical formulae. As for the methods used in our system in order to realize smoother handwriting input of mathematical formulae and other details about this interface, readers are referred to the paper [5].

3.2 Interface for users without sight

A trial version of InftyEditor is provided with a speech interface to read and edit mathematical documents.

As is well known, ordinary screen readers cannot correspond to math editors in which mathematical expressions are usually treated as images. On the other hand, InftyEditor treats all mathematical expressions as marked-up texts in XML. We, therefore, have a chance to access this editor with speech output.

Actually, we have already developed an interface for the InftyEditor to help visually disabled persons with accessing math expressions by means of speech in Japanese[3]. Since, in Japan, there is no definite or systematic method to convey content of math expressions accurately by speech, we firstly assign manners of aloud-reading in Japanese to all 712 symbols and math structures defined in the Infty Editor: such as a fraction, a subscript, a superscript, a radical, an integral

[3] One of the authors of the present paper with visual disability is using the trial version of the Japaneese speech interface of InftyEditor to read and edit scientific documents in his daily work

and so on. We adopted "95 Reader," the one of popular Japanese screen readers, as a speech output system to read aloud both of bodies of Japanese documents and control menus.

Next, we are currently trying to extend it to correspond to the English version of InftyEditor. In English-speaking countries, contrary to Japanese, it is usually defined properly how to read aloud mathematical expressions. In developing the interface, we are based on that method in principle and assign manners of aloud-reading to all necessary symbols and mathematical formulae structures. We use Microsoft speech API and Text-to-Speech engine as a speech output system for bodies of English documents. On the other hand, control menus are supposed to be read by a screen reader itself. Using this interface, one can accurately understand content of mathematical expressions by means of speech output only. For instance, a fraction is read aloud as follows:

"Frac a plus b over c frac-end."

That is, one can clearly realize where the beginning and the end of the fraction are located, what is the numerator and so on. Furthermore, not only one can access given mathematical expressions, but functions of writing and editing mathematical expressions are also available with a voice guide.

Combining this interface with the OCR technology for mathematical documents, it is expected that persons with visual disabilities become able to do freely reading and writing of scientific documents for themselves.

4 Output Formats

InftyReader outputs the recognition results in our original XML format, called 'KML'. The KML format includes the most detailed information obtained by the recognition results, such as block segmentation information (tag name, coordinates, etc.), line information (type of each line, indentation, centering, display math., equation number etc.), character information (candidates of recognition, coordinate, math/text attribute) and link information of the tree structure of mathematical expressions.

When the recognition result is loaded on InftyEditor's display, it is compressed into another XML format, called 'IML' omitting candidates and coordinates etc., information unnecessary to understand the content of the documents.

By the converter tool attached to Infty system, the XML files generated by InftyReader (or InftyEditor) can be converted into various formats of mathematical documents; LaTeX, MathML, HTML, and especially for users with visual disability, HR-TeX (Human Readable TeX), KAMS (Karlsruhe ASCII Mathematics Script) and Braille Codes, in UBC (Unified Braille Codes) for English texts and in Japanese Braille Codes for Japanese texts. The HR-TeX file is a text file in which mathematical expressions are written in simplified LaTeXnotations for convenience to be read by users with visual disability, omitting various commands and simbols unnecessary to understand the meaning. KAMS is a notation of mathematics developed in the Study Centre for Blind and Partially Sighted Students at the University of Karlsruhe.

There are several Braille notations which express mathematical formulae. Among them, we selected the Unified Braille Code (UBC) proposed by BANA (Braille Authority of North American) to output English mathematical documents. The main reason for this selection is the facility of the transcription. It needs lesser tasks to establish one-to-one correspondence between the printed form and UBC, for mathematical symbols and also in text word level for the Grade II codes, in UBC.

As for the Japanese texts, the Japanese formal Braille code defines the mathematical notations only for the high school level. The output of our system uses the notations extended by some volunteer but major transcriber groups to adapt it to the notation of the university level mathematics.

As it is described in [1], literal Japanese text is composed of a mixture of Chinese characters (called Kanji in Japanese), two sets of 46 Japanese syllabaries (Hira-gana and Kata-kana) and symbols. Before transcribing Japanese texts into Braille codes, it is necessary to transform the original text into syllabary text with correct pronunciation and proper punctuation. In our system, EXTRA which is a Japanese Braille transformation software proposed by Jun Ichikawa [6] is used to decode Kanji into syllabaries code.

Although the Braille emboss printer has a `columns` × `lines` restriction (ex. 12-42 characters per line × 10-28 lines per page), there is no automatic line fold function in our system at present. The main reason for it is that our system does not yet recognize the logical structure of the text such as chapters, sections, items, etc., or the mathematical description styles of theorems, definitions, etc.

5 Conclusion

The outline of the software called 'Infty' to recognize and transcribe printed mathematical documents into various digital data with accessibility for visually disabled people is presented. The system consists of three components of applications: an optical recognition system named 'InftyReader', an editor 'InftyEditor' and converting tool unit based on XML technology. InftyReader recognizes carefully scanned binary images in either 600 DPI or 400 DPI. It recognizes mathematical documents with complex formulae with high accuracy if the noise of scanned images are limited.

The recognized results by InftyReader can be edited easily by InftyEditor incorporated in the InftyReader package, including mathematical formulae, and the edited results can be transformed into various data formats: LaTeX, Human Readable TeX, Braille, etc. Infty Editor is featured with its handwriting interface to input mathematical formualae for users with sight and a speech interface to read and edit mathematical notations for users without sight.

Current version of InftyReader does not output the results of logical structure analysis of texts such as section-subsection structure, items, theorem descriptions etc. These subjects are left to future research work, but urgent since the logical structure analysis is inevitable to get correct output of Braille transcription.

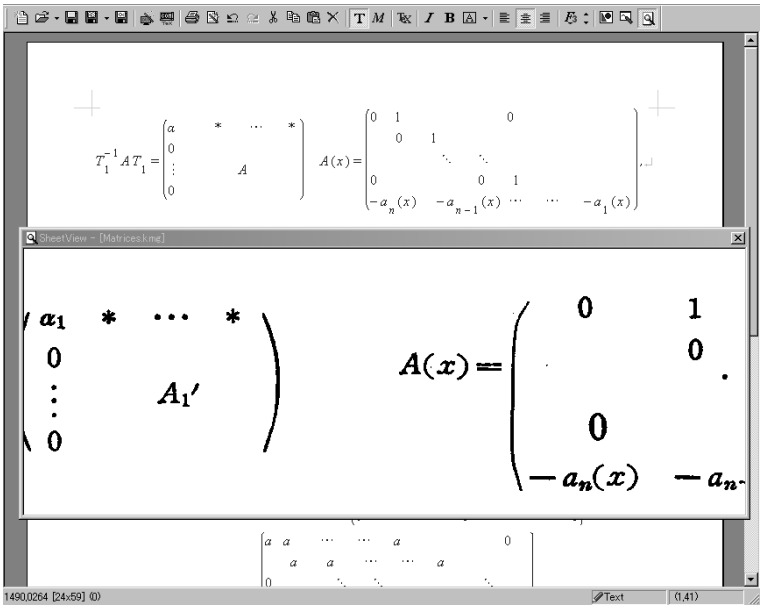

Fig. 3. Recognition Results of Matrices by InftyReader on InftyEditor's Display

To end this paper, we would like to mention the importance of the distinction of bold fonts in mathematical expressions omitted by the current version of InftyReader's output. It is important, since bold fonts are generally used in mathematical or scientific documents to distinguish vectors and scalar quantities.

References

1. R. Fukuda, N. Ohtake and M. Suzuki, "Optical Recognition and Braille Transcription of Mathematical Documents," *Proc. ICCHP*, 711–718, 2000.
2. Y. Eto and M. Suzuki, "Mathematical formula recognition using virtual link network," *Proc. ICDAR*, 762–767, 2001.
3. T. Kanahori and M. Suzuki, "A recognition method of matrices by using variable block pattern elements generating rectangular area," Graphics Recognition. Algorithms and Applications (Lecture Notes in Computer Science, 2390), Springer-Verlag, 2001.
4. M. Suzuki, F. Tamari, R. Fukuda, S. Uchida and T. Kanahori, "INFTY – An Integrated OCR System for Mathematical Documents," *Proc. DocEng*, 2003.
5. T. Kanahori, K. Tabata, W. Cong, F. Tamari and M. Suzuki, "On-Line Recognition of Mathematical Expressions Using Automatic Rewriting Method," *Proc. ICMI*, 394–401, (Lecture Notes in Computer Science, 1948), Springer-Verlag, 2000.
6. J. Ishikawa, *EXTRA for Windows ver. 1.0 users manual*, Amedia Co., Ltd., Tokyo, 2001.
7. `http://infty.math.kyushu-u.ac.jp`