/olume 7, Number 4, April 2013 ISSN 1881-803X

ICIC Express Letters

An International Journal of Research and Surveys

Indexed by *Ei Compendex (Elsevier)*
Scopus (Elsevier)
INSPEC (IET)

Published by ICIC International
http://www.ijicic.org/icicel.htm

ICIC Express Letters

ISSN 1881-803X

Printed in Japan

ICIC Express Letters
Volume 7, Number 4, April **2013**

ICIC International ©2013 ISSN 1881-803X
pp. 1431 1436

RESEARCH ON METHODS AND KEY TECHNOLOGIES OF MEANING EXTRACTION FROM MATHEMATICAL FORMULAS BASED ON MULTI-MODAL INFORMATION

Yao Liu and Ruijia Wang

Information Technology Support Center
Institute of Scientific and Technical Information of China
No. 15, Fuxing Road, Beijing 100038, P. R. China
liuy@istic.ac.cn

Received June 2012; accepted August 2012

Abstract. *We focus on building the description structure and system between image and semantics, and we will fully describe the multi-modal information, including the layout structural information, symbol syntax rules, formula syntactic rules and explanatory text, and realize the unity and fusion of the information in different types; we will also develop key technologies and realize the automatic semantic annotation of formulas. The methods proposed in this paper can reconstruct the mathematical formula and get the accurate calculational meaning of mathematical formula, and support the semantic analysis and advanced application of mathematical formula.*

Keywords: Mathematical formula, Formula recognition, Multi-modal, Natural language processing, Semantic annotations

1. **Research Significance.** Multi-modal is a new research field, and the research of multi-modal is generally focused on solving a certain problem with at least two modalities of information. There are a lot of mathematical formulas in scientific and technical literature, mathematical formula is a kind of multi-modal data, and it is the combination of image, signs and text features. The multi-modal information could complement the text, which is the main information in the literature, and helps the users to fully understand the knowledge in the scientific and technical literature. Therefore, it becomes one of the key problems to utilize the multi-modal information more effectively to help the computers recognize and understand the mathematical formula, and help the scientific research personnel to find related knowledge and understand the meaning and evolution rules of mathematical formula.

Many processing methods of mathematical formula structure analysis have been proposed over the years, including the formula decomposition methods using layout information (image features), and the formula understanding and description methods using syntax rules (symbol features), but many of the processing results are not satisfied, and developing the multi-modal information technology research based on the existing technology is the general trend. Therefore, we propose the mathematical formula meaning extraction theories and methods using multi-modal information processing technology based on the natural language processing technology.

2. Research Status and Problems Both Domestic and Overseas.

2.1. **Research status analysis both domestic and overseas.** The research on mathematical formula recognition has several stages: (1) Character recognition. (2) Structure analysis. (3) Combination of character recognition and structure analysis. (4) Local exploration, such as the combination of multi-modal fusion recognition and the character recognition, the combination of character recognition and text processing.

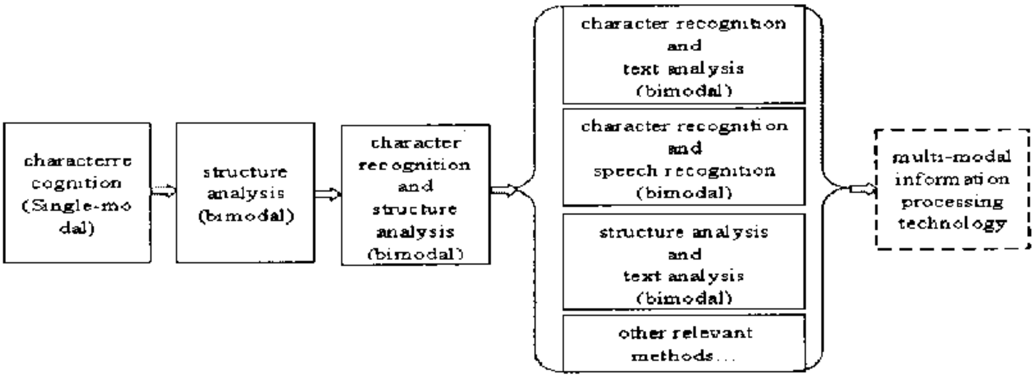
characterre cognition (Single-modal)
structure analysis (bimodal)
character recognition and structure analysis (bimodal)
character recognition and text analysis (bimodal)
character recognition and speech recognition (bimodal)
structure analysis and text analysis (bimodal)
other relevant methods...
multi-modal information processing technology

FIGURE 1. The stages and trend of mathematical formula recognition

From the perspective of modality information processing, the research on mathematical formula recognition has three stages: (1) Single-modal processing, corresponding to the character recognition. (2) Simple bimodal processing, which combines the character recognition and structure analysis. (3) Diversified bimodal processing. In this stage of mathematical formula recognition research, some bimodal processing methods have been proposed besides the simple bimodal processing methods, such as the combination of character recognition and text analysis, the combination of character recognition and speech recognition, and the combination of structure analysis and text analysis.

The stages and trend of mathematical formula recognition are shown in Figure 1.

In the rest of this chapter, these research stages will be introduced respectively.

2.1.1. *Character recognition stage.* Character recognition is also called OCR, which is "Optical Character Recognition". As a hot domain in research, it has a history of thirty years [1,2], and provides a solid base for the specific symbol recognition in mathematical formula. The main difference between and OCR is that the Formula character recognition technology separates the formula from the text in the article, analyzes the formula structure and reconstructs it. Other processing steps of formula character recognition are similar with those steps in OCR system, especially the formula symbol recognition.

2.1.2. *Structure recognition stage.* Anderson addressed the problem of formula recognition in his doctoral thesis for the first time in 1968 [3], but he only sketchy described the concept of mathematical formula recognition, and did not gave the complete theory or test data. In the following decade, mathematical formula recognition technology developed slowly because of the complexity of formula, and the limitations of image recognition and computer technique.

2.1.3. *Character recognition and structure recognition stage.* Blostein and Grbavec [4] defined the mathematical formula recognition for the first time, they divided the mathematical formula recognition into two stages: symbol recognition and structure analysis. Belaid and Haton [5] used syntactic analysis to analyze the formula more concise; Dimitriadis and Coronado [6] designed a mathematical editor; Chan and Yeund [7] designed an online handwritten mathematical expressions recognition system utilizing the structural analysis and syntactic analysis.

2.1.4. *Application of new technologies.* Inoue et al. [8] designed a system based on the original OCR system to process Japanese documents. Kacem, Belaid and Benahmed [9] proposed a new solution based on the previous work, and received satisfied results. Chaudhuri and Garain [10] utilized the variance calculation in statistics to detect the mathematical formula in a line of text.

2.2. **Existing problems.** The two-dimensional space structure of the mathematical formulas makes it quite difficult to locate and extract formulas. We can accurately recognize every symbol in the formula, but we cannot always recognize or extract the whole formula.

These difficulties mentioned above increase the difficulty for mathematical formula recognition, the existing problems of mathematical formula recognition are as follows: (1) There is no mature theory system. There is no mature theory system which can fully satisfy the actual demands in this research field, most existing algorithms either cannot get satisfied recognition rate, or have limited formula types and symbols that can be recognized. (2) The multi-modal information is not fully utilized. The research on multi-modal is generally focused on solving a certain problem with at least two modalities of information. Mathematical formula is a kind of multi-modal data, it is the combination of image, signs and text features. The multi-modal information could complement the text, which is the main information in the literature, and helps the users to fully understand the knowledge in the scientific and technical literature.

3. Main Research Contents.

3.1. **Theoretical and method research of mathematical formula multi-modal semantic analysis.** Analyze the main fields, progress and development direction of multi-modal research both domestic and overseas, and emphasize the integration of multi-subjects.

3.2. **Inherent rule and semantic correlation of mathematical formula multi-modal features.** Discuss the relationship between multi-modal and semantics, find the semantic correlation between mathematical formula features of different modalities, and construct a structured semantic description system which focuses on the entities, relationships and events in scientific and technical literature, in order to generate semantic representation of the mathematical formulas.

3.3. **Single mode information analysis and feature extraction.** Develop integrate methods to analyze and extract semantic multi-modal information, and generate effective analysis and extraction of the multi-modal information in mathematical formulas.

3.4. **Multi-modal mathematical formula semantic feature extraction and expression.** Find the statistical relationship between different multi-modal mathematical formula semantic features based on natural language processing technologies, construct co-occurrence matrix of formula multi-modal features to generate isomorphic subspaces with different data types and reflect the relevance, and finally generate the expression of relationship between semantic features of different modalities.

3.5. **Multi-modal fusion and expression model construction based on context.** Develop algorithm which is suitable for fusion and collaborative analysis of mathematical formula multi-modal information, and generate the fusion of mathematical formula multi-modal semantic features and develop theories and the methods to extract features of mathematical formula multi-modal information.

4. **Technical Route and Key Technology.** Our research combines every stage of formula recognition based on natural language processing, and focuses on the effective fusion theory and methods of text and other information in different processing stages. This research will develop related key technologies and realize the effective extraction and indexing of formula meanings. The research programme and overall technical route are shown in Figure 2, the key research contents are shown in bold and have original creativity, other research contents mainly utilize the existing research results.

The purpose of mathematical formula recognition is to convert formula images into editable formulas in text form by image processing, character segmentation and structure

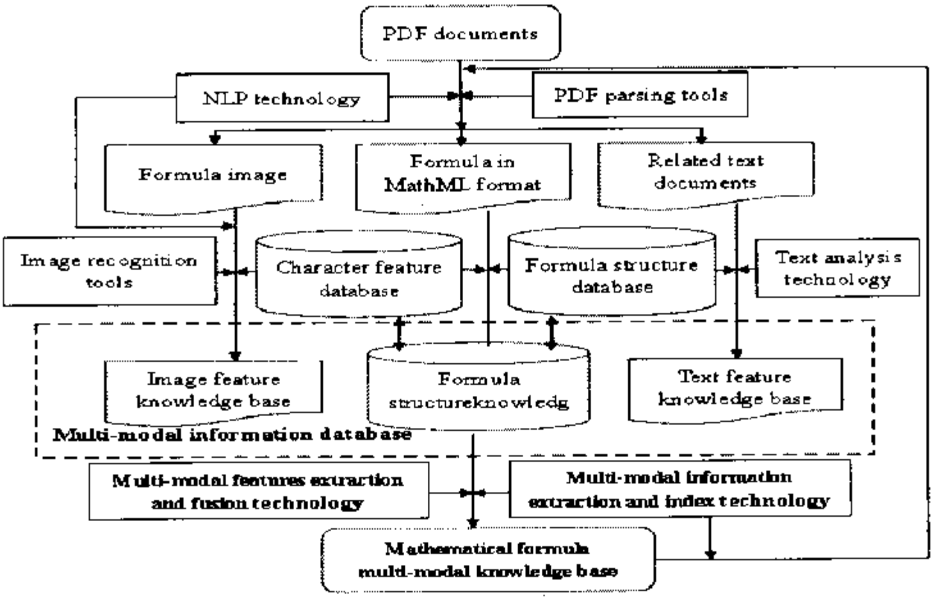
PDF documents
NLP technology
PDF parsing tools
Formula image
Formula in MathML format
Related text documents
Image recognition tools
Character feature database
Formula structure database
Text analysis technology
Image feature knowledge base
Formula structureknowledg
Text feature knowledge base
Multi-modal information database
Multi-modal features extraction and fusion technology
Multi-modal information extraction and index technology
Mathematical formula muld-modal knowledge base

FIGURE 2. Research programme and overall technical route

reconstruction. Generally speaking, there are five steps in the conversion from formula image to formula in text form as follows:

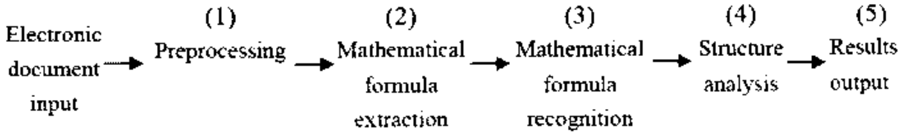
Electronic document input
(1) Preprocessing
(2) Mathematical formula extraction
(3) Mathematical formula recognition
(4) Structure analysis
(5) Results output

FIGURE 3. Mathematical formula recognition steps in image

We take the PDF documents as the example to conduct the research:

(1) The construction methods and realization of mathematical formula comprehensive knowledge base

Our research mainly develops the construction methods and technologies of formula structure knowledge base, image feature knowledge base and text feature knowledge base, and completes the knowledge construction on a certain scale (to support the experiment).

A. Formula structure knowledge base

Ideas: formula structure + character feature + formula explanation text feature → formula structure knowledge base

Solution: Develop the methods and technologies of formula recognition and extraction based on the features, extract the markup language format based on XML in the Web and the editable formula based on the ASCII linear string format, and realize the automatic construction of formula structure knowledge base, or help the construction of this knowledge base.

B. Image feature knowledge base

Ideas: character database + formula explanation text feature → image feature knowledge base

Solution: The research on character recognition is relatively mature and the resource (character database) acquirement is convenient, therefore, based on the existing research, we focus on the research on the associated relationships between character features and formula explanation text, especially the relationships and rules between the location information and the related text, in order to construct the image feature knowledge which is suitable for formula semantic indexing.

C. Text feature knowledge base

Ideas: character feature + location information + formula structure information + text feature → text feature knowledge base

Solution: Our research utilizes the text recognition technology and the features in this research field to construct the formula text feature knowledge base, we focus on exploring the associated relationships between the explanation text and the formula character meanings, in order to construct the text feature knowledge base.

(2) Development of multi-modal information extraction and indexing technology

Ideas: Construct the multi-modal features fusion model and realize the core algorithm.

Solution: Aim for the semantic information acquirement, this research combines the construct process of mathematical formula comprehensive knowledge base to construct a description structure and system between image and semantics. We will fully describe the multi-modal information, including the layout structural information, symbol syntax rules, formula syntactic rules and explanatory text, and realize the unity and fusion of the information in different types.

(3) Development of multi-modal information extraction and indexing technology

Ideas: Develop related extraction and indexing tools and rudimentary implement the automatic indexing or auxiliary indexing for mathematical semantic.

Solution: Based on the comprehensive knowledge base, this research combines the knowledge construction technology to develop related tools and realize the automatic indexing or auxiliary indexing for formula meanings in PDF documents, this research will also realize the supplement and development of the content in formula comprehensive knowledge base.

5. **Conclusions.** We utilize the multi-modal information to acquire and represent the mathematical formula semantic information for the first time, and research on the multimodal semantic information representation and relations of mathematical formulas, in order to improve the semantic representation of mathematical formula contents and provide the foundation for knowledge service.

Aimed at the polymorphism of mathematical formula, this research explores the complementary of multi-modal features, construct a system to present the semantic multi-modal features of mathematical formula and realize the unity and fusion of the information in different types.

In this paper, we construct a description structure and system between image and semantics, develop key technologies and realize the automatic semantic annotation of formulas. The methods proposed in this paper can reconstruct the mathematical formula and get the accurate calculational meaning of mathematical formula, and support the semantic analysis and advanced application of mathematical formula.

Acknowledgment. This work is partially supported by National Key Project of Scientific and Technical Supporting Programs No. 2011BAH10B04, National Social Science Fund No. 12BTQ006, The authors also gratefully acknowledge the helpful comments and suggestions of the reviewers, which have improved the presentation.

REFERENCES

[1] S. Mori, C. Y. Suen and K. Yamamoto, Historical review of OCR research and development, *Proc. of IEEE*, vol.80, no.7, pp.1029-1058, 1992.

[2] C. C. Tappert, C. Y. Suen and T. Wakahara, The state of the art in online handwriting recognition, *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol.12, no.8, pp.787-808, 1990.

[3] R. H. Anderson, *Syntax-Directed Recognition of Hand-Printed Two-Dimensional Equations*, Harvard University, Boston, 1968.

[4] D. Blostein and A. Grbavec, Recognition of mathematical notation, in *Handbook of Character Recognition and Document Image Analysis*, H. Binke and P. Wang (eds.), Singapore, World Scientific, 1997.

[5] A. Belaid and J. P. Haton, A syntactic approach for handwritten mathematical formula recognition, *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol.6, no.1, pp.105-111, 1984.

[6] Y. A. Dimitriadis and J. L. Coronado, Towards an AIU'based mathematical editor, that uses on-line handwritten symbol recognition, *Patter Recognition*, vol.28, no.6, pp.807-822, 1995.

[7] A.-F. Chan and D.-Y. Yeund, An efficient syntactic approach t structural analysis of online handwritten mathematical expressions, *Information Sciences*, vol.33, no.3, pp.375-384, 2000.

[8] K. Inoue, R. Miyazaki and M. Suzuki, Optical recognition of printed mathematical documents, *Proc. of ATCM'98*, pp.280-289, 1998.

[9] A. Kacem, A. Belaid and M. Benahmed, Automatic segmentation of mathematical documents, *Proc. of ACIDCA00*, Monastir-Tunisia, pp.86-91, 2000.

[10] B. B. Chaudhuri and U. Garain, An approach for recognition and interpretation of mathematical expressions in and interpretation of mathematical expressions in printed document, *Pattern Analysis & Applications*, vol.11, no.3, pp.120-131, 2000.