EXTRAFOR : automatic EXTRAction of mathematical FORmulas

Afef Kacem, Abdel Belaid, Mohamed Ben Ahmed

EXTRAFOR : automatic EXTRAction of mathematical FORmulas

[1]A. Kacem, [2]A. Belaïd and [3]M. Ben Ahmed
[1] ENSI-RIADI, afef.kacem@isg.rnu.tn
[2]LORIA-CNRS, abdel.belaid@loria.fr
[3]ENSI-RIADI, mohamed.benahmed@serst.rnrt.tn

Abstract

We present a method for automatic extraction of mathematical formulas from images of documents without character recognition. Formula extraction is first done by location of its most significant symbols, then extension to adjoining symbols using contextual rules until delimitation of the whole formula space. Mathematical symbols labelling is realised from models created at the learning stage using fuzzy logic. From the experiments, we found that the average rate of primary labelling of mathematical symbols is about 95.3%.The obtained results have demonstrated the applicability of our system since 90% of mathematical formulas are well extracted from documents printed with high quality.

1. Introduction

A document may consist of various kinds of components, such as text, images, graphics, and mathematical formulas. Most systems proposed to analyse images of printed documents cannot handle all kinds of components. In fact, these components, having different structures and typographies, need to be separated in order to be analysed more efficiently by dedicated systems. This paper intends to present a system which will separate formulas from other components of document.

Mathematical formulas are involved in scientific documents, either as isolated expressions, or embedded directly into a text-line. Thus, the first step in mathematics recognition is to identify where expressions are located on the page. Recently several researchers have proposed algorithms for recognition of mathematical expressions [1-11]. They showed some sound recognition results. But most of the work we survey assume that recognition system begins with an isolated mathematical expression. An exception is Lee and Wang [14], who present a method for extracting both embedded and isolated mathematical expressions in a text document. Text lines are labelled as isolated expressions based both on internal properties and on having increased white space above and below them. The remaining text lines consist of a mixture of pure text and text with embedded expressions. No details of this process are given, except that it is done "according to some basic expressions forms".

So several problems still exist in these recognition systems, which include how to extract automatically mathematical expressions from a document, how to improve the recognition rate of multifont characters, and how to correct the recognition errors in a mathematical expression. In this paper, we simplify the problem by only extracting mathematical formulas from documents to be able to read, parse and re-use them in other applications.

In a mathematical formula, characters and symbols can be arranged as a complex two-dimensional structure, possibly of different characters and symbol sizes. They differ greatly from text since a line of text is one-dimensional and discrete : characters are placed one after another on the same line when symbols in formulas may be under, above, on the right and far, included in another, etc, with continuous distances. This makes its extraction process more complicated even when all the individual characters and symbols can be recognised correctly.

To restore planar structure to formulas, two solutions are often proposed : recognition of characters, then restructuring or else labelling then recognition. The first solution assumes that optical reader has succeeded to segment formulas and is able to provide location of each character. The second solution simplifies the work, since it segments formula in characters before presenting them individually to an optical reader system. This method avoids OCR segmentation procedures which are too general. Noticing the lack of success of the first method, we have experimented the second using adaptive segmentation of text. The idea is to do labelling at several steps : extraction of lines, then location of isolated and embedded formulas.

This paper is organised as follows : in the next section, we present the overview of our system. We then describe our current approach to identifying mathematical symbols,

analysing and extending their context in a way to extract formulas from text. Afterwards, we present and discuss some experimental results which are then followed by some concluding remarks, We finally, give some conclusions and prospects.

2. EXTRAFOR system

Fig.1. shows the overview of EXTRAFOR, a system that we developed to extract mathematical formulas automatically from images of documents.

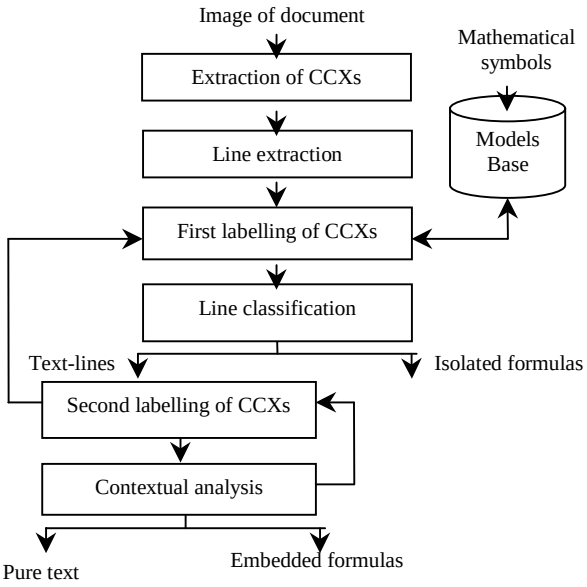


Fig.1. Overview of EXTRAFOR system

2.1. Extraction of CCXs

At this stage, the document is scanned, its image is straightened and its connected components (CCXs) are extracted. These CCXs constitute the basic datum from which our system starts its analysis. Each CCX is described by co-ordinates of upper left (X_{min}, Y_{min}) and lower right (X_{max}, Y_{max}) corners of its bounding box and the number of its black pixels (NBP). Let W and H be respectively the width and the height of CCX, then the following parameters are determined : aspect ratio, R=W/H, area A=W*H and density D=NBP/A.

After extraction of CCXs, it is convenient to restrict ulterior processing to those susceptible to be symbols of formulas which could improve precision and speed of their extraction. The filtering of CCXs is based on their area and aspect ratio to avoid noise, diacritical, punctuation signs, graphs, horizontal and vertical separators.

2.2. Extraction of lines

This step consists in grouping, into lines, horizontally adjacent CCXs. First, CCXs are sorted by ascending Y_{min}. Then, lines co-ordinates (X_{minl}, Y_{minl}, X_{maxl}, Y_{maxl}) are updated from CCXs having a common intersection of their heights. Once a set of CCXs is associated to a particular line, we can sort them by ascending X_{min}.

After that, it is convenient to assemble CCXs of not linear formulas as numerators and denominator lines of fractions as well as limit expressions of summations or products. In fact, their CCXs may be separated after line extraction step as seen in Fig.2. Line fusion phase needs results of CCXs primary labelling and their proximity study.

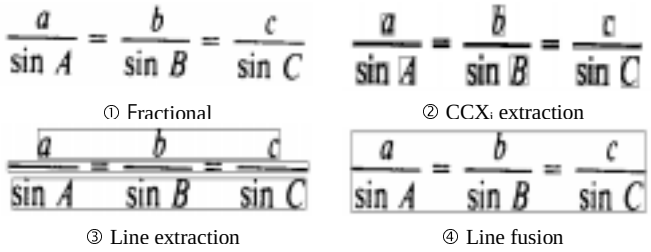


Fig.2. Line extraction and fusion

2.3. First labelling of CCXs

This is about locating CCXs representing mathematical symbols. A label is assigned to each component according to the role it could play in formula composition. We will distinguish between Functional Symbols (FS: summations or product signs), Integral (IS) and Radical Signs (RS), Horizontal Fraction Bars (HFB), Small and Vertical Great Delimiters (SD, VGD), Binary Operators (BO: signs of subtraction), Subscripts and Superscripts (SUB, SUP). To classify these different types of mathematical symbols, we have thought of topographical (position according the central band), morphological (aspect ratio) and typographical (area and density) classification of their CCXs. (See Table.1. and Fig.3.)

Topography	Morphology	Typography	Symbols
Overflowing	Squared, Great	Enlarged, Normal	FS
Overflowing, Ascending	Large	Enlarged	RS
Overflowing	Very extensive	Enlarged	IS
Centred	Very lengthened	Dense, Very dense, Enlarged, Normal	HFB
Overflowing	Very extensive	Enlarged	VGD
Ascending	Extensive	Normal	SD
Centred	Lengthned	Very dense, reduced	BO
Descending, Deep			SUB
Ascending, High			SUP

Table.1.Symbols classification

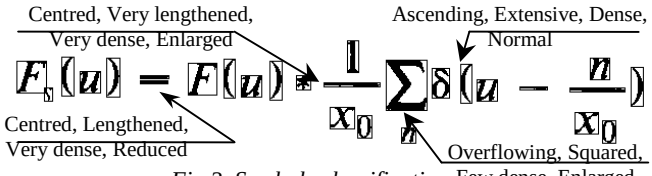


Fig.3. Symbols classification

2.3.1. Mathematical symbols training

For an effective classification, the system must analyse the greatest possible number of symbols deduced from different mathematical documents to extract ranges of aspect ratio, density and area from their CCXs. These intervals being representatives of symbol type.

To have an assessed idea of similarity that can exist between different typographies and morphologies of the same symbol, we have taken a sample of mathematical symbols with different fonts, areas and styles. For each instance of symbol, values of aspect ratios, areas and densities are calculated, observed and only lower and upper bounds (LB, UB) are considered. We have studied 175 FS, 57 RS, 69 IS, 92 HFB, 116 VGD, 205 SD and 140 BO. The following results were obtained (See Table.2.).

Symbols	LB(R)	UB(R)	LB(A)	UB(A)	LB(D)	UB(D)
FS	0.26	1.64	198	3900	0.23	0.48
RS	1	7.94	1435	47850	0.05	0.2
IS	0.16	0.69	660	8832	0.10	0.29
HFB	8	87.71	138	6336	0.14	1
VGD	0.05	0.26	345	4840	0.14	0.70
SD	0.06	0.41	116	990	0.22	0.93
BO	4	13.5	42	125	0.62	1

Table.2. : Training results

2.3.2. Fuzzy identification of CCXs

To identify mathematical symbols, we have introduced degrees of membership to different symbol classes and constitute corresponding histograms. The abscissa of histograms represents all classes : the whole measured values shared in regular width intervals. The ordinate is the relative frequency: the number of measurements belonging to a class, divided by the total number of measurements. The ordinate can be considered as the membership degree to a class. It varies between 0 and 1. The generated histograms have to be as representative as possible, so the closest to a continuous function.

To identify a mathematical symbol given its CCX, values of each parameters P={R,D,A}are calculated. By referring to histograms of each type of symbol, we each time keep the membership degree of that CCX to a type of symbol: MS according to one parameter noted $\mu_{MS,P}(CCX)$. We then keep, for each type of symbol, the minimal membership degree of that CCX according to its aspect ratio, density and area. We finally take their maximal value. Thus, the membership degree of the CCX to the type of symbol MS is defined by: $\mu_{MS}(CCX)=Max(Min(\mu_{MS,R}(CCX),\mu_{MS,A}(CCX),\mu_{MS,D}(CCX))=Max(\mu_{FS}(CCX), \mu_{IS}(CCX), \mu_{RS}(CCX), \mu_{HFB}(CCX), \mu_{VGD}(CCX), \mu_{SD}(CCX), \mu_{BO}(CCX))$. Table.3. presents an example of fuzzy identification of one small delimiter where R= 0.27, D= 0.32 and A= 418.

MS	$\mu_{MS,R}$(CCX)	$\mu_{MS,D}$(CCX)	$\mu_{MS,A}$(CCX)	μ_{MS}(CCX)
FS	0.04	0.16	0.22	0.04
RS	0	0	0	0
IS	0.40	0	0	0
HFB	0	0.12	0.76	0
VGD	0	0.36	0.20	0
SD	0.35	0.44	0.49	0.35
BO	0	0	0	0
			μ_{MS} (CCX)	0.35
			MS	SD

Table.3.: Fuzzy identification of a small delimiter

The obtained label corresponds to the true type of symbol although there could be confusion with class of functional symbols. But, it is clear that $\mu_{SD}(CCX)$ is superior than $\mu_{FS}(CCX)$.

2.4. Lines classification

Once lines are extracted and their CCXs are labelled, isolated formulas could be located. Generally, the height of lines containing isolated formulas is superior than the average height of lines. Besides, they are often centred that is the distances that separate them from the right and left margins are almost equal. Based on theses characteristics, isolated formulas could be extracted from images of documents, which restrict next stages to text-lines which could contain embedded formulas. Elsewhere, we decide to abandon processing of formulas which are very linear since they can be recognised by OCR systems.

2.5. Second labelling of CCXs

This concerns CCXs of text-lines. It is a finer labelling of CCXs, belonging to the same line, in which we have considered their position from central band to solve certain ambiguities observed at their primary labelling. In fact, the topographical classification of CCXs could distinguish between functional symbols and characters or digits similarly between integral symbols and oblique fraction bar, since integral and functional symbols are overflowing while characters, digits and oblique fractions bars are not. In addition, with this classification, subscripts and superscripts of formulas could be detected since their CCXs are generally deepen or higher. For those having descending or ascending components, we have considered two other features which are : the relative size : X=RS/LS (RS: Right component Size, LS: Left component Size) and the relative position : Y=D/LH (D: Distance between the top of the right component and the bottom of the left component).

2.6. Contextual analysis

At this stage, we tried to delimit embedded formulas. We have considered mathematical formula as a set of sub-expressions that could be spread to the left or the right. Initially, these sub-expressions are symbols of formula. Then, by ascending successive fusion, these symbols will

include their operands and other neighbouring expressions in a way to separate formulas from pure text. We present here some heuristics rules we used to extend context of mathematical symbols.

-R1: If a subscript or superscript is found, then it is grouped with its closest neighbour.

-R2: If a radical symbol encloses others components, then their fusion is a formula.

-R3: If an horizontal fraction bar is found, then its numerators and denominators are joined to it.

-R4: If a functional or an integral symbol is found, then its lower and upper bounds in addition of the first component of its sub expression are joined to it.

-R5: Each CCX enclosed inside a pair of vertical great delimiters should form a formula.

-R6: A sign of subtraction, a subscript or a superscript, a functional, a radical, an integral symbol or a horizontal fraction bar or else a reduced number of characters enclosed inside a pair of small delimiters should form a formula.

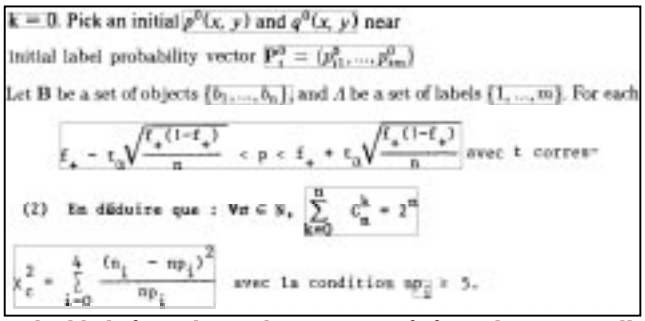
-R7: If a binary operator is found, then its left and right operands are joined to it.

-R8: Two horizontal adjacent formulas constitute one formula

-R9: Two formulas, separated by a reduced number of CCXs compose one formula.

3. Experimental results

We carried our experimental formula extraction for a variety of mathematical documents. Images of documents are scanned at a resolution of 300dpi. In the experiments, we trained our system using 854 mathematical symbols and tested 460 symbols and about 50 formulas. Fig.4 shows the results obtained after extraction of some



embedded formulas. About 90% of formulas are well extracted form mathematical documents. The main errors are due to confusion of subscripts and superscripts with diacritic signs.

Fig.4. : Examples of formulas extraction

The main difficulties concern functional and integral formulas inserted in a text where we must detect the presence of a specific character like the 'd' of an integral formula or one character of the lower limit of a functional formula.

4. Conclusion and prospects

In this paper, we have developed a method to extract formulas automatically from images of mathematical documents without using OCR system. We have introduced fuzzy logic at CCXs labelling which has been useful to identify symbols and consequently to delimit formulas by a contextual analysis of their CCXs. Thus, we have been able to separate them from other components of the document.

Our prospects will include: dealing with more complex alignment symbols of formulas and confirming the efficiency and the performance of our method for a large data base of mathematical formulas.

References

[1] ANDERSON R.H.,« Two-Dimensional Mathematical Notation », in **Syntactic Pattern Recognition Applications**, K.S. Fu, Ed. Springer Verlag, NewYork , 1977, pp.147-177.

[2] BELAID Abdelwaheb, **HATON** Jean-Paul,« A syntactic ApProc.h for Handwritten Mathematical Formula Recognition », in **IEEE Trans. PAMI**, vol 6. N°1, janvier 1984, pp. 105-111.

[3] GRBAVEC Ann, **BLOSTEIN** Dorothea,« Recognition of mathematical notation", Handbook of character recognition and document image analysis, world scientific publishing company, 1997, pp. 557-582.

[4] HASHIM M.Twaakyondo, **MASAYKI** Okamoto,« Structure Analysis and Recognition of Mathematical Expressions », in **ICDAR'95**, Canada, 1995, pp.430-437.

[5] HSI-Jian Lee, **MIN**-Chou Lee,« Understanding Mathematical Expression in a Printed Document », in **ICDAR'93**, Japan,1993, pp.502-505.

[6] JAEKYU Ha, **HARALICK** Robert M., **IHSIN** T. Phillips, "Understanding mathematical expressions from document images", in **ICDAR'95**, Canada, 1995, pp. 956-959.

[7] LAVRIOLLE Stéphane**, POTTIER** Loïc,"Optical formula recognition", ICDAR'97, Canada, pp 357-361, 1997.

[8] MASAYUKI Okamoto, **AKIRA** Miyazawa,« An experimental Implementation of Document Recognition System for Papers Containing Mathematical Expressions », in **Structured Document Image Analysis,** Springer, verlag, pp. 36-53, 1992.

[9] XUEJUN Zhao**, XINYU** Liu**, SHENGLING** Zheng**, BOACHANG** Pan**, TANG** Yuan Y.,"On line recognition handwritten mathematical symbols", **ICDAR'97**, Allemagne, 1997, pp. 645-648.

[10] WANG Zi-xiong, **FAURE** Claudie, "Structural analysis of mathematical expressions", **9th ICPR**, Washington, 1988, pp. 32-34.

[11] CHANG S. K.,« A Method for the Structural Analysis of 2-D Mathematical Expressions », in **Information Sciences**, Vol. 2, N°3, pp.253-272, 1970.

[12] HSI-Jian Lee, **JIUMN**-Shine Wang,«Design of mathematical expression recognition system », in **ICDAR'95**, Japan,1995, pp.1084-1087.