Identification of Embedded Mathematical Expressions in Scanned Documents

Utpal Garain  
B. B. Chaudhuri  
Computer Vision & Pattern Recognition Unit  
Indian Statistical Institute  
203, B. T. Road, Kolkata-108, India  
{utpal, bbc}@isical.ac.in

A. Ray Chaudhuri  
Dept. of Computer Science & Engg.  
Jadavpur University  
Kolkata 700 032, INDIA  
chaudhuri_ar@yahoo.co.in

Abstract

Efficient extraction of mathematical expressions is considered as an important pre-processing step to apply existing OCR systems to convert scientific papers into their electronic format. In this correspondence, a technique for extracting embedded (or in-line) expressions has been presented. The proposed method for expression extraction initially invokes an existing OCR to recognize the input document. Several features including word n-grams (a statistical analysis of a corpus of scientific documents reveals that the word level n-gram profile for sentences containing embedded expressions is quite different from that of the sentences without any expression) are computed on sentence level to spot sentences containing expressions. Expression zones are pinpointed by exploiting OCR inability to handle expressions and by using some common typographical aspects followed in typing mathematical expressions. Experimental results on a considerable size of dataset show high efficiency of the proposed technique.

1. Introduction

It has been observed that the exiting OCR systems perform poorly for mathematical expressions contained in scientific documents. This makes the conversion of scientific papers from printed to electronic form a difficult task. In such documents, mathematical expressions may appear in two modes (see fig. 1): (i) embedded (also, called in-line expressions) i.e. mixed with normal text and (ii) displayed (also, called isolated expressions) i.e. typed in separate line.

The presence of expressions disturbs an existing OCR system not yet trained for expression recognition and therefore, identification and extraction of expressions may result in an efficient conversion process for scientific paper documents. Such a framework will allow an existing OCR engine to process the normal text portion as usual whereas the extracted expressions can be processed by an OCR specially designed for expression recognition [1, 2].

In our earlier study [3], we presented an expression extraction technique that provided high accuracy for identifying displayed expressions but showed less accuracy for embedded ones. This paper presents an alternative algorithm to efficiently extract embedded expressions contained in scientific documents.

Figure 1. A sample page containing embedded as well as displayed expressions.

2. Review of previous studies

Among the existing studies dealing with expression extraction, the technique proposed by Lee and Wang [4] locates embedded expressions by recognizing characters in a text line and converting them to a stream of tokens. A token is decided to belong to an embedded expression according to some basic expression forms which considers presence of special mathematical symbols, super-scripting, or matrix structures. Symbols that are adjacent to the above tokens are heuristically attached to form an embedded expression. Fateman [5] presented a 3-pass algorithm that initially recognizes all connected components and separates them into two bags, math and text. The text bag contains all Ro-
man letters, italic numbers and the math bag includes punctuation signs, special symbols, italic letters, Roman digits, and other marks like lines, dots, etc. Next, components in the math bag are grouped into zones according to their proximity. Symbols that are left ungrouped and appeared to be too far from other math symbols are moved to the text bag. Symbols in the text bag are similarly joined up into groups according to proximity. Text words (hopefully include words like “sin”, etc.) that are relatively isolated from other text but within any previously identified math zone are moved to the math bag. Segmentation result is finally reviewed by human assistance to correct errors, if any.

The method proposed by Inoue et. al. [6] isolates expressions contained in Japanese scientific document by assuming that the OCR recognizes Japanese characters with high confidence whereas expression symbols are either rejected or recognized (rather misrecognized) with low confidence. In another approach, Toumiit et. al. [7] locate embedded expressions by finding special symbols like “=”, “+”, “<”, “>”, etc. and some specific context propagation from these symbols is done. For example, for parenthesis and brackets, symbols between them are checked; for horizontal bars, symbols above and below them are investigated; etc.

Later on, Kacem et. al. [8] proposed a 2-pass (global and local segmentation) scheme where expressions are initially separated from text lines using a primary labeling which uses fuzzy logic based model to identify some mathematical symbols. A secondary labeling uses some heuristics to reinforce the results of the primary labeling and locates super/sup-scripts inside the text. An evaluation strategy has been presented to judge the expression extraction technique and a success rate of about 93% has been reported on a combined test set of 300 displayed and embedded expressions.

Recently, Chowdhury et. al. [9] proposed a recognition-free approach that exploits the usual spatial distribution of the black pixels in math zones. Experimental results show that the method works well for segmenting displayed expression but gives only 68.08% accuracy for extraction of embedded expression. In another recognition-free technique reported by Jin et. al. [10], embedded expressions are extracted based on the detection of two-dimensional structures. However, the authors of [9] and [10] concluded that the extraction of embedded expressions is quite difficult without doing character recognition.

3. Our Proposed Approach

In our approach, we have adopted a technique that is quite different from the previous ones. Our study is started with a corpus [11] of 400 scanned pages of scientific documents containing 3101 embedded expressions showing large variety in constituent symbols and their structural arrangement. Next, two well-known commercial OCR systems are invoked to recognize these pages. Recognition results for sentences with and without expressions are separated for further investigation. The following observations are important in this context.

- Sentences without expressions are recognized with almost no error.
- Also, high recognition accuracy is obtained for normal text words in sentences with expressions.
- Some of the expression symbols (e.g. Roman letters, digits, symbols like “+”, “-”, “=”, sometimes punctuation marks, etc.) are often recognized properly. However, for majority of these symbols, the OCRs, on recognition, associate suspicion marks with them to indicate that either these symbols in isolation do not form any valid word (e.g. isolated letters, letter with scripts, words like “sin”, “log”, etc.) or to reveal poor OCR confidence during their recognition (sometimes for italic or bold characters).
- Other expression symbols (mostly Greek letters, majority of mathematical operators, special symbols, etc.) are either rejected (signaled by a special symbol) or misrecognized with a suspicion mark.
- If word level n-grams are computed for two categories of sentences (with and without expressions) then such an n-gram based category profiles markedly differ from one another. Let $C_E$ denotes the category of sentences containing embedded expressions.

In our approach, we have captured these observations and formulated a knowledge base that helps to identify and extract embedded expressions. For a given document, the proposed method, at first, invokes an existing OCR to recognize the page content in its capability. Next, for each sentence word n-grams are computed and its category is determined against the pre-computed statistics. In fact, instead of categorizing a sentence, a probability that the sentence belongs to the category $C_E$ is computed.

Within each sentence, words (a sequence of characters delimited by a space) for which OCR shows less recognition confidence are checked for their type style (i.e. italic, bold, etc.). Type style of a word is checked at the image domain by following the technique described in [12]. To obtain the binary image of a word, image level bounding box information (i.e. the four corner coordinates) is made available for each word along with its recognition result. Let $Y_{max}$ denotes the lower-most y-coordinate of a symbol and the variance of $Y_{max}$ of the symbols is estimated to verify the geometric alignment of the symbols within the word.

A word in a sentence is labeled as mathematical by looking at five parameters (i) the probability that the sentence containing the word belongs to the category $C_E$, (ii) confidence level of the OCR while recognizing the constituent symbols of the word, (iii) type style of the word, (iv) inter-character spacing within the word, and (v) variance of $Y_{max}$. 


Let \( t_{A1} \) and \( t_{A2} \) denote the access times of \( M_1 \) and \( M_2 \), respectively, relative to the CPU.

Let \( tA1 \) and \( tA2 \) denote the access time of \( M_1 \) and \( M_2 \), respectively, relative to the CPU.

**Figure 2. Finding of embedded expressions.**

of the constituent symbols. If more than one word in a sentence are labeled as mathematical, then they are grouped together according to their positional proximity to form embedded expressions.

Let us explain the approach considering Fig. 2. Fig. 2(a) shows an image of a line containing four expression fragments. The OCR output for this sentence is shown in fig. 2(b). Words for which the OCR shows less confidence are underlined in fig. 2(b). At first, the sentence shows its inclination towards the category \( C_F \) mainly because of the presence of the words “let” and “denote”. Word-level investigation reveals that confidence scores for recognizing “tA1”, “tA2”, “M1” and “M2” are less than the average confidence for other words. Moreover, italic style is detected for some of the constituent symbols of these words, e.g. type style for “t” and “A” are found italics when the word “tA1” is investigated. The variance of \( \gammaX \) of the symbols ‘t’, ‘A’ and ‘t’ and the average inter-character spacing are computed and compared with the average values observed in the words for which acceptable recognition scores were obtained. A larger variance and a lesser average inter-character gap indicate the word “tA1” to be a mathematical one. Following this approach all the four expression fragments are identified as embedded expressions. For identifying function words like “sin”, “log”, “exp”, etc., a finite automata consisting of 12 final states corresponding to 32 commonly used function words is maintained in the system.

**4. Experimental Results and Discussion**

Experimental has been carried out on 400 scanned pages taken from various science books, journals, conference proceedings, etc. These pages contain 1084 sentences with 3101 expression fragments embedded in them. A sentence is said to have one or more embedded expressions if it would need the use of math mode had the sentence been prepared using \( \LaTeX \). For each page (say, DOCxxx.tif), embedded expressions contained in the page are truthed in a separate file called DOCxxx.emb. Along with other information [11] each expression is tagged with its bounding box corner (top-left and bottom-right) coordinates.

Extraction of expressions is evaluated by against the bounding box information available in the groundtruthed data. One of the four cases will occur during this checking: **Case-I.** An extraction is correct if the bounding box corresponding to the extracted zone finds a match in the groundtruth. **Case-II.** Extracted zone shows a bounding box that partially matches the groundtruth. **Case-III.** An expression could not be extracted (i.e. totally missed) and **Case-IV.** An extracted zone does not contain any mathematics at all (i.e. false identification). Occurrences of the first three cases are shown in fig. 3 with examples. A score (for example, \( \alpha \), \( \beta \), and \( \gamma \) for case-I, II and III, respectively) is associated with each type of cases. Let \( T_1 \) be the total number of expressions properly extracted, \( T_2 \) be the number for which partial extraction is done, \( T_3 \) be the number of expressions, which are missed (i.e. not extracted) and \( T_4 \) be the number of zones that do not contain any expressions (false identification). The extraction efficiency \( (E) \) for an input page is computed as

\[
E = \frac{\alpha T_1 + \sum_{i=1}^{3} \beta_i + \gamma T_3}{\alpha T} - \left( 1 - e^{-\delta T} \right)
\]

where, \( T = T_1 + T_2 + T_3 \). In our evaluation strategy, \( \alpha \) and \( \gamma \) are set to 1 and 0, respectively. \( \beta_i \) is computed as \( \beta_i = 1 - \frac{C_{a_i} - C_{e_i}}{C_{e_i}} \) where \( C_e \) is the number of components found within the i-th extracted zone and \( C_a \) is the number of components actually present in the zone. The value of \( C_a \) is obtained from the groundtruth where MathML presentation tags are available for each expression. The last term i.e. induces penalty due to false identification. In our system, \( \delta \) is set to 1, however, an empirically chosen value for \( \delta \) may penalize false identifications is a more judicious manner.

If a document \( D_i \) shows an extraction accuracy of \( E_i \), then an average efficiency \( (E_{avg}) \) is computed in a dataset of \( N \) documents as follows:

\[
E_{avg} = \frac{1}{N} \sum_{i=1}^{N} E_i
\]

Following the equation (1), extraction efficiencies for each of the 400 pages are computed separately. For the page portion in fig. 1, extraction results are shown in fig. 3. Out of 11 expressions, 8 are properly extracted, 2 are missed and 1 is partially extracted. For partially extracted expression, the extracted zone contains 2 out of 3 components giving \( \beta_i = 1 - \frac{2}{3} = 0.67 \). Therefore, extraction efficiency \( (E) \) equals to 0.788 \([1*(1+0.67+0.67)*2]/1*11\] is obtained. The equation (2) combines page-level efficiency scores to give an overall measure for a given set of documents. For our case, a value 0.963 for \( E_{avg} \) has been obtained for extracting 3101 expression embedded in 400 pages. Details of extraction results are given in Table 1.

\[\text{http://www.w3.org/Math/}\]
Table 1. Summary of Extraction Results

<table>
<thead>
<tr>
<th>Nature of Extraction</th>
<th>Accuracy</th>
<th>Remarks</th>
</tr>
</thead>
<tbody>
<tr>
<td>Case-I (Perfect)</td>
<td>2904/3101 (93.65%)</td>
<td>Detection of Perfect, Partial and Missed is done against the groundtruthed data.</td>
</tr>
<tr>
<td>Case-II (Partial)</td>
<td>169/3101 (5.45%)</td>
<td></td>
</tr>
<tr>
<td>Case-III (Missed)</td>
<td>28/3101 (0.9%)</td>
<td></td>
</tr>
<tr>
<td>Case-IV (False)</td>
<td>62</td>
<td>In this case, some text portions are wrongly extracted as expressions.</td>
</tr>
</tbody>
</table>

Major extraction errors occur for some of the short expressions containing only 2 or 3 symbols and for documents with excessive degradation due to aging, etc. where large number of broken and merged characters appear and disturb the extraction process. The performance measure presented in equation (1) also considers the false extractions along with the accepted and rejected extractions. In our experiment, analysis of test results shows that to extract 3101 expression zones, 62 zones fall under this false extraction category that induces a penalty of 0.0199 (negative) in the overall performance measure.

5. Conclusion

A technique for the extraction of embedded expression contained in scientific documents is presented. The approach presented in this paper uses features some of which were not explored in the previous studies. Experiment involving a test set of considerable size shows encouraging results. Integration of the proposed approach with one of the existing OCR systems and evaluation of its performance in recognizing scientific documents will be considered in a future work.

References