11

OCR of Printed Mathematical Expressions

Utpal Garain and Bidyut B. Chaudhuri

11.1 Introduction

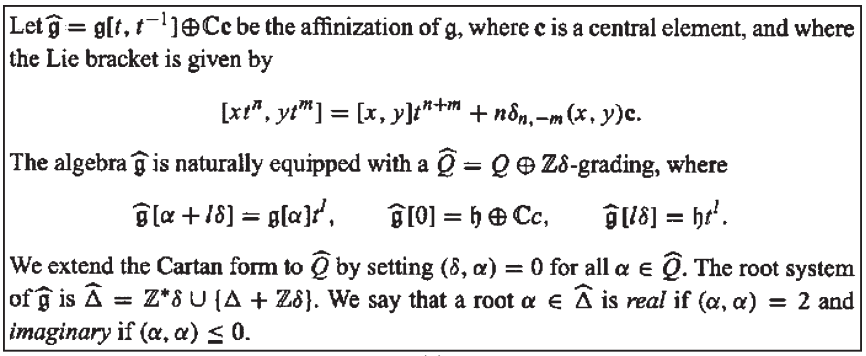
Automatic recognition of mathematical expressions (hereafter, referred as expressions) is one of the challenging pattern recognition problems of significant practical importance. Such a recognition task is required while converting printed scientific documents into electronic form or to aid the visually impaired persons in reading scientific documents.

This chapter discusses the key issues involved in the development of a system for recognition of printed documents containing expressions. In fact, studies on this topic date back to 1960s when Anderson [1] proposed a syntax-directed scheme for recognition of hand-printed expressions. Several studies have been reported so far and surveyed in [4, 5, 13].

However, the present discussion starts with showing limitations of existing optical character recognition (OCR) systems in converting scientific papers into corresponding electronic form. Figure 11.1 demonstrates one such example obtained from a popular OCR system, namely ABBYY Fine Reader 6.0.[1] The limitation arises because such a system fails to recognize expressions that often appear in scientific documents.

Instead of developing new OCR systems for scientific documents, we put emphasis on upgrading the existing ones by additional processing modules for expressions in documents. Since the presence of expressions disturbs typical OCR system (not trained for expression recognition), the identification and extraction of expression zones, therefore, could be the first step in this module. It permits an existing OCR engine to process the normal text portion as usual, whereas the extracted expressions can be tackled by a system specially designed for expression recognition.

[1] www.abbyy.com.



Let $\widehat{\mathfrak{g}} = \mathfrak{g}[t, t^{-1}] \oplus \mathbb{C}\mathbf{c}$ be the affinization of $\mathfrak{g}$, where $\mathbf{c}$ is a central element, and where the Lie bracket is given by

$$[xt^n, yt^m] = [x, y]t^{n+m} + n\delta_{n,-m}(x, y)\mathbf{c}.$$

The algebra $\widehat{\mathfrak{g}}$ is naturally equipped with a $\widehat{Q} = Q \oplus \mathbb{Z}\delta$-grading, where

$$\widehat{\mathfrak{g}}[\alpha + l\delta] = \mathfrak{g}[\alpha]t^l, \qquad \widehat{\mathfrak{g}}[0] = \mathfrak{h} \oplus \mathbb{C}c, \qquad \widehat{\mathfrak{g}}[l\delta] = \mathfrak{h}t^l.$$

We extend the Cartan form to $\widehat{Q}$ by setting $(\delta, \alpha) = 0$ for all $\alpha \in \widehat{Q}$. The root system of $\widehat{\mathfrak{g}}$ is $\widehat{\Delta} = \mathbb{Z}^*\delta \cup \{\Delta + \mathbb{Z}\delta\}$. We say that a root $\alpha \in \widehat{\Delta}$ is *real* if $(\alpha, \alpha) = 2$ and *imaginary* if $(\alpha, \alpha) \leq 0$.

(a)

Lefg = $Q[I, t^1]$ j ©Cc be the affinization of g, where c is a central element, and where the Lie bracket is given by

$$[xt^n, yt^m] = (X_{>}y]t^{n+m} + n\delta_{n,-m}(x, y)C.$$

The algebra $\hat{g}$ is naturally equipped with a $Q = Q \odot Z5$-grading, where

$$f[a + IS] = Q[a]t', \qquad f[0] = fj \odot Cc, \qquad ?[/\$] = Ij^{*'}.$$

We extend the Cartan form to Q by setting (5, a) = 0 for all a e Q. The root system off is A = Z*5 U {A + Z<5}. We say that a root « 6 A is rea/ if $(a,$ a) = 2 and *imaginary* if (a, a) < 0.

(b)

Fig. 11.1. OCR output of scientific documents: an example (a) image (b) recognition results.

Mathematical expressions typically appear in documents, either as (a) displayed (isolated) expressions or (b) expressions embedded into (i.e. mixed with) the text lines. As far as automatic identification of expressions is concerned, displayed and embedded expressions impose different level of complexities. The displayed ones are typed in separate lines and exhibit several image-level features that distinguish them from normal text lines. On the other hand, embedded expressions are generally small expression fragments, which are difficult to isolate from the text portion mixed with expressions. These issues have been discussed in Section 11.3.

Once expressions in a document are identified, recognition of them involves two major components namely (i) symbol recognition and (ii) structure interpretation. Symbol recognition is difficult because a large character set (roman letters, Arabic digits, Greek letters, Operator symbols, etc.) with a variety of typefaces (regular, italic, bold), and a large number of different font sizes may be used to generate the expressions. Moreover, certain symbols (e.g. *integration*, *summation*, *product*, *brackets*, etc.) are elastic in nature and have a wide range of possible scales.