Contents lists available at ScienceDirect

Pattern Recognition

journal homepage: www.elsevier.com/locate/pr

An integrated grammar-based approach for mathematical expression recognition

Francisco Álvaro*, Joan-Andreu Sánchez, José-Miguel Benedí

Pattern Recognition and Human Language Technologies Research Center, Universitat Politècnica de València, 46022 Valencia Spain

ARTICLE INFO

Article history:
Received 7 April 2015
Received in revised form
29 July 2015
Accepted 5 September 2015

Keywords:
Mathematical expression recognition
Probabilistic parsing
Handwriting recognition

ABSTRACT

Automatic recognition of mathematical expressions is a challenging pattern recognition problem since there are many ambiguities at different levels. On the one hand, the recognition of the symbols of the mathematical expression. On the other hand, the detection of the two-dimensional structure that relates the symbols and represents the math expression. These problems are closely related since symbol recognition is influenced by the structure of the expression, while the structure strongly depends on the symbols that are recognized. For these reasons, we present an integrated approach that combines several stochastic sources of information and is able to globally determine the most likely expression. This way, symbol segmentation, symbol recognition and structural analysis are simultaneously optimized. In this paper we define the statistical framework of a model based on two-dimensional grammars and its associated parsing algorithm. Since the search space is too large, restrictions are introduced for making the search feasible. We have developed a system that implements this approach and we report results on the large public dataset of the CROHME international competition. This approach significantly outperforms other proposals and was awarded best system using only the training dataset of the competition.

1. Introduction

Mathematical notation constitutes an essential source of information in many fields. Recognizing mathematical expressions is an important problem in scientific document recognition or the acceptance of mathematical expressions in human-based interfaces [1]. Lately, other applications like accessibility for disabled people and, especially, information retrieval are receiving more attention [2].

In this pattern recognition problem we usually distinguish between online and offline expressions. Offline formulas are represented as images and can be printed or handwritten. Meanwhile, an online expression is encoded as a sequence of points in space, and therefore includes time information. Handwritten expressions generally present more variability than printed expressions, while online data normally yield better results than offline expressions because they contain additional information [3]. This paper is focused on automatic recognition of online handwritten mathematical expressions.

The recognition of math notation is traditionally divided into three problems [1,2]: symbol segmentation, symbol recognition and structural analysis. In the literature, there are two main approaches: sequential (decoupled) solutions and integrated solutions.

Sequential approaches tend to first look for the best segmentation of the input expression into math symbols. The analysis of the structure is then carried out on that best symbol segmentation [4]. This type of solution is not able to solve errors made during the first recognition stage if hard decisions are taken.

Integrated approaches set out to use the global information of the mathematical expression to obtain the final structure of the formula as a whole [5,6]. The symbol segmentation and the symbol recognition are then obtained as a byproduct of the global optimization. These approaches seem more appropriate because obtaining the best symbol segmentation as a byproduct depends on the structure of the expression, and vice versa.

In this paper we present a novel integrated approach for the recognition of online handwritten mathematical expressions. This approach is inspired in the *holistic* approach used in *off-line* Handwritten Text Recognition (HTR) [7]. In the holistic HTR approach, the recognition process is modeled in several perception levels integrated in a unique model: optical models (Hidden Markov Models [8] or Recurrent Neural Networks [9]) are used for modeling characters in the low level; finite-state models are used for modeling words in the middle level [8]; and n-gram models are used for language modeling in the high level.

We follow an analogous approach in this paper with different perception levels integrated in a unique model. We describe a

* Corresponding author.
E-mail addresses: falvaro@prhlt.upv.es (F. Álvaro), jandreu@prhlt.upv.es (J.-A. Sánchez), jbenedi@prhlt.upv.es (J.-M. Benedí).

http://dx.doi.org/10.1016/j.patcog.2015.09.013
0031-3203/© 2015 Elsevier Ltd. All rights reserved.

statistical framework based on two-dimensional grammars for the highest level since they constitute a natural and powerful model for dealing with structured problems. Mathematical symbols are used in the middle level that are modeled as sets of strokes, and strokes as primitives in the lowest level. It is important to remark that other input units smaller than strokes could be considered in the lowest level (like subsets of consecutive input points) but the principal ideas of the approach described in this paper would be the same.

We define the associated parsing algorithm that globally determines the most likely math expression based on several sources of information. The aspects that allow us to obtain the segmentation and recognition of symbols as a byproduct are explained in detail.

In the integrated approach, the search space becomes enormous and therefore we also present techniques based on spatial and geometric information for effectively reducing the search space. We impose restrictions based on the distance between strokes, and during the structure analysis we impose restrictions following the idea of hierarchical clustering algorithms. Finally we tackle the estimation of all the probability distributions.

The system that implements this proposal was awarded best system using the training set of the recent CROHME international competition [10], and has been released as open-source software. Here we report results on the large public dataset of this competition.

The paper is organized as follows. First, related work is presented in Section 2. The statistical framework of our novel grammar-based approach is described in Section 3. This framework derives two different models: the symbol likelihood that is described in Section 4, and the structural probability that is defined in Section 5. The parsing algorithm associated with this statistical framework is given in Section 6, and the experimentation carried out, along with analysis and discussion of the results, is reported in Section 7. Finally, conclusions and future work are presented in Section 8.

2. Related work

The problem of automatic mathematical expression recognition has been studied for decades [11]. Many approaches have been proposed [12,4,13], but unfortunately most of them cannot be properly compared due to the lack of public datasets or standard metrics.

The Infty corpus [14] was released several years ago as a great resource of printed math expressions. More recently the MathBrush dataset [15] provided another important resource for handwritten math expression recognition. Over the last few years, the rapid growth of tactile devices and human-based interfaces has brought more attention to handwriting recognition solutions. With the recent editions of the CROHME competition [10], and the development of a set of performance evaluation metrics [16,17], mathematical expression recognition has become a very active research field. In last three editions of the CROHME competition, systems were submitted from nine different countries.

Different approaches have been presented for math expression recognition. Zanibbi and Blostein [4] recognized an expression as a tree, and proposed a system based on a sequence of tree transformations. Eto and Suzuki [18] developed a model for printed math expression recognition that computed the minimum spanning tree of a network representation of the expression. Shi et al. [19,20] presented a system where symbol segmentation and recognition were tackled simultaneously based on graphs. They then generated several symbol candidates for the best segmentation, and the recognized expression was computed in the final structural analysis [21].

Given the well-known structure of mathematical notation, many approaches are based on grammars because they constitute a natural way to model this problem. In fact, the first proposals on math expression recognition were grammar-based [11,12]. Since then, different studies have been developed using different types of grammars. For instance, Chan and Yeung [22] used definite clause grammars, the Lavirotte and Pottier [23] model was based on graph grammars, Yamamoto et al. [24] presented a system using Probabilistic Context-Free Grammars (PCFG), and MacLean and Labahn [13] developed an approach using relational grammars and fuzzy sets. In this paper we will focus on models based on PCFG.

Proposals based on PCFG use grammars to model the structure of the expression, but the recognition systems are different. Garain and Chaudhuri [25] proposed a system that combines online and offline information in the structural analysis. First, they created online hypotheses based on determining baselines in the input expression, and then offline hypotheses using recursive horizontal and vertical splits. Finally they used a context-free grammar to guide the process of merging the hypotheses. Yamamoto et al. [24] presented a version of the CYK algorithm for parsing bidimensional PCFG (2D-PCFG) with the restriction that symbols and relations must follow the writing order. They defined probability functions based on a region representation called “hidden writing area”. Průša and Hlaváč [26] described a system for offline recognition using 2D context-free grammars. Their proposal was penalty-based such that weights were associated with regions and syntactic rules. The model proposed by Awal et al. [5] considers several segmentation hypotheses based on spatial information, and the symbol classifier has a rejection class in order to avoid incorrect segmentations.

In this paper we present a formal model that is grounded in two studies. First, the system for math expression recognition based on parsing 2D-PCFG presented in [6]. That model tackled symbol segmentation by computing the connected components of the input strokes and merging them using productions of the grammar (e.g. an equal sign is a line below another line). However, this strategy required consideration of additional classes for symbol composition (e.g. an *i* without the dot or linked letters in functions) and finding proper spatial relations for their combination. Moreover, it could not account for touching symbols or symbols with segmentations that have not been taken into account with specific productions in the grammar (for instance, broken symbols like π in Fig. 5). Thus, segmentation was not a hidden variable but depended on design decisions of the grammar and symbol composition. Also, in [6] the estimation of the 2D-PCFG was not tackled.

Second, in order to overcome the problems of this segmentation methodology, we integrated a stroke-based approach similar to [19,20] into the parsing process of 2D-PCFG. The solution presented in [19,20] combined several stochastic sources of information on symbol segmentation, symbol recognition and symbol relationships in order to determine the best overall segmentation and recognition. However, it had the restriction that symbols must be written with consecutive strokes in time and structural analysis was performed as a decoupled step.

As a result, in this paper we develop a statistical framework for mathematical expression recognition which main contributions with regard to [6,19,20] are as follows: (i) a fully integrated approach based on 2D-PCFG using strokes as primitives with no time order assumptions. Our proposal integrates several stochastic information sources in order to globally determine the most likely mathematical expression. In this advanced framework, segmentation becomes a hidden variable; (ii) we deal with the estimation of all the probabilistic sources of information and the reduction of the search space; (iii) the framework is able to deal with all

Please cite this article as: F. Álvaro, et al., An integrated grammar-based approach for mathematical expression recognition, Pattern Recognition (2015), http://dx.doi.org/10.1016/j.patcog.2015.09.013