Pattern Recognition Letters 35 (2014) 58–67

Contents lists available at SciVerse ScienceDirect

Pattern Recognition Letters

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

Recognition of on-line handwritten mathematical expressions using 2D stochastic context-free grammars and hidden Markov models

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

Departamento de Sistemas Informáticos y Computación, Universitat Politècnica de València, Valencia, Spain

ARTICLE INFO

Article history:
Available online 6 October 2012

Keywords:
Mathematical expression recognition
Handwriting recognition
Stochastic parsing
Symbol recognition
Structural analysis
Spatial relations

ABSTRACT

This paper describes a formal model for the recognition of on-line handwritten mathematical expressions using 2D stochastic context-free grammars and hidden Markov models. Hidden Markov models are used to recognize mathematical symbols, and a stochastic context-free grammar is used to model the relation between these symbols. This formal model makes possible to use classic algorithms for parsing and stochastic estimation. In this way, first, the model is able to capture many of variability phenomena that appear in on-line handwritten mathematical expressions during the training process. And second, the parsing process can make decisions taking into account only stochastic information, and avoiding heuristic decisions. The proposed model participated in a contest of mathematical expression recognition and it obtained the best results at different levels.

1. Introduction

An essential part of information in science documents and many other fields is mathematical notation. Introducing mathematical expressions into computers and handling them usually requires special notation like LaTeX or MathML. However, in recent times there has been a great increasing of pen-based interfaces and tactile devices that allow users to provide handwritten data as input. This is a more natural way of introducing mathematical notation and it requires, in turn, developing systems that are able to recognize them. Recognition systems for mathematical expressions depend on the application (Blostein and Grbavec, 1997): on-line recognition systems for handwritten mathematical expressions, and off-line recognition systems for handwritten or printed mathematical expressions. In on-line recognition, the system input is usually a set of strokes that have geometric and temporal information. The system can take profit of temporal information of mathematical expressions that is not present in off-line recognition. This paper will be focused in on-line recognition of handwritten mathematical expressions.

Handwritten mathematical expression recognition can be divided into two major steps (Chan and Yeung, 2000b): symbol recognition and structural analysis. Symbol recognition involves segmentation of the input strokes into mathematical characters and symbol classification of these hypotheses. Structural analysis deals with finding out the structure of the expression according to the symbols arrangement. Several methods have been proposed to solve these problems jointly and separately. If the symbol recognition and structural analysis problems are performed jointly, then the segmentation can be considered as a hidden variable. If the symbol recognition and structural analysis problems are carried out separately, then the segmentation is done explicitly. The second approach is usually computational less expensive but it is prone to segmentation errors. The first approach is computational more expensive and efficient A* search strategies must be defined (Rhee and Kim, 2009).

Several proposals have been studied for printed symbol recognition based on Support Vector Machines (Malon et al., 2008), Nearest Neighbor techniques or Hidden Markov Models (HMM) (Álvaro and Sánchez, 2010). A survey of methods is provided in (Chan and Yeung, 2000b) for on-line recognition. The most usual representation techniques for on-line recognition are based on template or structural matching and HMM. Currently the error rate of mathematical symbols is about 5% for the best known classification techniques (Garain and Chaudhuri, 2004; Luo et al., 2008; Shi et al., 2007). A very low error rate in mathematical symbol segmentation is crucial because it largely contributes to a reduction in the recognition of the full expression.

Structural analysis of mathematical expression has been carried out by considering syntactic models that are able to represent spatial relations between the mathematical symbols. Trees are the mostly accepted models and they have been considered in (Chan and Yeung, 2000a; Chou, 1989; Průša and Hlaváč, 2007; Yamamoto et al., 2006; Zanibbi et al., 2002) to represent these 2D spatial relations in on-line recognition. Minimum spanning trees have also been considered (Eto and Suzuki, 2001). These tree structures are

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

0167-8655/$ - see front matter © 2012 Elsevier B.V. All rights reserved.
http://dx.doi.org/10.1016/j.patrec.2012.09.023

obtained through a parsing process. Evaluation and comparison of the structural analysis in mathematical expression recognition has hitherto been difficult because: first, most of the proposals used private data, and second, there has been a lack of standard performance evaluation measures (Awal et al., 2010; Lapointe and Blostein, 2009). Lately, some public large corpora have been developed (MacLean et al., 2011; Suzuki et al., 2003) and also new metrics have been proposed (Álvaro et al., 2011a; Sain et al., 2010; Zanibbi et al., 2011). Recently a Competition on Recognition of Online Handwritten Mathematical Expressions[1] (CROHME 2011) has been proposed (Mouchére et al., 2011). This competition tried to define an experimental setup that allowed the comparison of different systems.

The contribution of this paper is twofold. First, from a theoretical point of view, this paper takes a step forward in mathematical expression recognition by introducing a stochastic formal model based on 2D stochastic context free grammars (SCFG). The main contribution of this paper is an adequately formalized model that allowed us to deal appropriately with the two fundamental problems associated to this kind of models, that is, the learning problem, and the interpretation problem. The learning problem deals with the learning of the structural and stochastic part of the model from data. This paper introduces the learning of the main stochastic distributions of the model. The interpretation problem deals with obtaining the structure of the mathematical expression. The algorithm for parsing mathematical expressions according to the proposed model is also defined in this paper. It is important to remark that a correct formalization of the model allows us to use classical probabilistic estimation algorithms for learning and efficient algorithms for stochastic parsing.

Second, from a more applied point of view, this paper contributes to mathematical symbol recognition and spatial relations classification as follows. HMM have been widely used for symbol recognition, in this work we applied two well-known sets of features used in handwritten text recognition (Toselli et al., 2004, 2007) that have not been yet tested in mathematical symbol recognition. Regarding spatial relations classification, we extended the work presented in (Aly et al., 2009) for printed mathematical expressions. We propose adding new features to account for variability in the handwritten case, as well as to recognize spatial relations that were not considered in that paper. We also employed a different method for computing the geometric features.

A system for handwritten mathematical expression recognition based on parsing 2D-SCFG is completely detailed in this paper. It is composed by hybrid HMM as mathematical symbol classifier and spatial relations are determined according to geometric features. This model is able to solve jointly the problems involved in mathematical expression recognition: symbol recognition and structural analysis. The described system participated in the CROHME 2011 competition previously mentioned and it obtained the best results.

The remainder of the paper is organized as follows. First, a review of related works is given in Section 2. Then, an overview of the on-line handwritten mathematical expressions recognition system is presented in Section 3. The formal statistical framework based on a two-dimensional extension of SCFG is described in Section 4. The mathematical symbol recognition process is explained is Section 5 and the spatial relations classification is detailed in Section 6. Finally, Section 7 presents a set of experiments performed to validate this approach using a public database of mathematical expressions and also the results obtained in the CROHME 2011 competition are reported. Conclusions and future work are presented in Section 8.

[1] Competition held in the *International Conference on Document Analysis and Recognition* (ICDAR 2011). http://www.isical.ac.in/∼crohme/.

2. Related work

There are several papers with detailed surveys on on-line recognition of mathematical expressions (Chan and Yeung, 2000b; Garain and Chaudhuri, 2004; Vuong et al., 2008). In this section we will briefly describe those papers that have influenced at most the approach of this paper.

There are two main steps in mathematical expression recognition: symbol segmentation and recognition and structural analysis of the recognized symbols. On-line handwritten mathematical expressions are usually represented as a set of strokes. The segmentation problem consists on grouping properly these strokes to produce mathematical symbols. The segmentation problem has been tackled by computing connected components in (Průša and Hlaváč, 2007) or applying the projection profile cutting method in (Okamoto and Miao, 1991), and the obtained strokes are then classified. The primary unit representation in mathematical expression recognition can be mathematical symbols (Winkler, 1996) or strokes. When using strokes as primary unit representation, the segmentation into strokes can be explicitly used (Zanibbi et al., 2002; Garain and Chaudhuri, 2004; Yamamoto et al., 2006; Vuong et al., 2008; Rhee and Kim, 2009) or not (Shi et al., 2007; Luo et al., 2008). The use of explicit segmentation into strokes introduces the problem of grouping them to compose symbols, and it makes the search difficult. Given that currently there are powerful symbol segmentation techniques (Luo et al., 2008), in this paper, the primary unit representation was mathematical symbols.

Recognition of handwritten mathematical symbols is carried out using classical techniques. Several methods have been proposed to solve this problem, such as HMM (Winkler, 1996; Kosmala et al., 1999; Hu and Zanibbi, 2011), Neural Networks (Thammano and Rugkunchon, 2006), Elastic Matching (Chan and Yeung, 1998; Vuong et al., 2010) or Support Vector Machines (Keshari and Watt, 2007). Furthermore, some of these proposals combine on-line and off-line information to perform hybrid classification and improving recognition results (Winkler, 1996; Keshari and Watt, 2007). An important advantage of using HMM is that they may not require an explicit segmentation for training (Shi et al., 2007; Luo et al., 2008). In this work we have used HMM and we have combined on-line and off-line characteristics with HMM.

Several approaches have tackled the structural analysis problem by using formal grammars. Chou (1989) proposed to use SCFG in order to recognize printed mathematical expressions. Other proposals have been presented using definite clause grammars (Chan and Yeung, 2001) or graph grammars (Lavirotte and Pottier, 1998). Yamamoto et al. (2006) presented a statistical formulation for on-line parsing of handwritten mathematical expressions based on the Cocke–Younger–Kasami (CYK) algorithm. Our proposal is similar to the approach based on two-dimensional SCFGs that was described in (Průša and Hlaváč, 2007). Průša and Hlaváč used weights associated to regions that contained symbols and then these regions were combined according to a penalty function based on region weights and syntactic rule weights. Our proposal uses stochastic distributions associated to regions in a similar way to Yamamoto et al. (2006), but the stochastic distributions associated to the model were different in our proposal. The stochastic definition of our model allowed us to use classical parsing and estimation algorithms for SCFG (Lari and Young, 1990).

3. System overview

This section presents an overview of the developed system, and the following sections provide a detailed description of the most important parts of the model.